

METODOLOGÍA HÍBRIDA NO SUPERVISADA BASADA EN ALGORITMOS
GENÉTICOS Y *FUZZY C-MEANS* PARA LA IDENTIFICACIÓN DE PÉRDIDAS DE
ENERGÍA EN TRANSFORMADORES DE DISTRIBUCIÓN

HAMILTON JAVIER PELÁEZ ALZATE

Código: 4520468

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS: ELÉCTRICA, ELECTRÓNICA, FÍSICA Y
CIENCIAS DE LA COMPUTACIÓN
PROGRAMA DE MAESTRIA EN INGENIERÍA ELÉCTRICA
PEREIRA
2017

METODOLOGÍA HÍBRIDA NO SUPERVISADA BASADA EN ALGORITMOS
GENÉTICOS Y *FUZZY C-MEANS* PARA LA IDENTIFICACIÓN DE PÉRDIDAS DE
ENERGÍA EN TRANSFORMADORES DE DISTRIBUCIÓN

HAMILTON JAVIER PELÁEZ ALZATE

Código: 4520468

Proyecto de grado presentado como requisito para optar el título de magister en
Ingeniería Eléctrica.

Director
Ph.D Oscar Gómez Carmona

UNIVERSIDAD TECNOLÓGICA DE PEREIRA
FACULTAD DE INGENIERÍAS: ELÉCTRICA, ELECTRÓNICA, FÍSICA Y
CIENCIAS DE LA COMPUTACIÓN
PROGRAMA DE MAESTRIA EN INGENIERÍA ELÉCTRICA
PEREIRA
2017

AGRADECIMIENTOS

A Dios por permitirme culminar este proyecto y ser mi guía permanente

A mi esposa por su constante apoyo y comprensión

A toda mi familia por ser mi fuente de inspiración y motivación

A mis colegas y amigos por el ánimo recibido

A EDEQ por el suministro de la información para esta investigación

Al ingeniero Oscar Gómez por la orientación y acompañamiento en este trabajo

A la UTP por ser ese polo de desarrollo tan importante para la región

ÍNDICE GENERAL

Pág.

ÍNDICE GENERAL	4
1. PROBLEMA DE INVESTIGACIÓN	9
1.1. PLANTEAMIENTO DEL PROBLEMA	9
1.2. JUSTIFICACIÓN	10
1.3. ANTECEDENTES	11
1.4. OBJETIVOS	17
1.4.1. Objetivo general.....	17
1.4.2. Objetivos específicos	17
1.5. ESTRUCTURA DEL DOCUMENTO	18
2. MARCO TEÓRICO	19
2.1. LA MACROMEDICIÓN Y EL BALANCE DE ENERGÍA	19
2.2. EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS (KDD).....	21
2.2.1. La minería de datos	22
2.2.2. Algoritmos de agrupamiento en tareas de <i>Clustering</i>	22
2.2.2.1. Algoritmo de <i>Clustering Fuzzy C-Means</i>	25
2.2.2.2. El método jerárquico divisivo <i>Bisecting k-Means</i>	28
2.2.3. Normalización de los datos para un análisis de <i>Clustering</i>	29
2.2.3.1. Normalización <i>Z-score</i>	30
2.2.3.2. Normalización <i>Min-Max</i>	31
2.2.4. Criterio de agrupamiento basado en medidas de disimilaridad.....	31
2.2.4.1. Distancia Euclidiana	32
2.2.4.2. Coeficiente de correlación de Pearson	33
2.2.5. Índices de validación de <i>Clustering</i>	34
2.2.5.1. Índice Xie-Beni.....	35
2.2.5.2. Índice Davis-Bouldin	36
2.3. AGRUPACIÓN DE SERIES DE TIEMPO	36
2.4. ALGORITMOS GENÉTICOS	38
3. METODOLOGÍA PROPUESTA	42
3.1. ADQUISICIÓN E INTEGRACIÓN DE LA INFORMACIÓN.....	42
3.2. PRE-PROCESAMIENTO Y NORMALIZACIÓN DE LOS DATOS.....	45

3.3.	PROCESAMIENTO DE LOS DATOS	47
3.3.1.	Agrupación mediante Algoritmo Genético (AG)	47
3.3.1.1.	Generación población inicial	48
3.3.1.2.	Criterio de agrupación	50
3.3.1.3.	Evaluación función <i>Fitness</i>	50
3.3.1.4.	Operadores genéticos	51
3.3.1.5.	Criterio de parada	54
3.3.2.	Agrupación mediante metodología híbrida FCM-AG	55
3.3.3.	Método genético divisivo <i>Bisecting Fuzzy C-Means</i> (GBCM)	56
4.	RESULTADOS Y CONCLUSIONES	60
4.1.	AGRUPACIÓN CON AG	60
4.2.	AGRUPACIÓN CON FCM-AG	61
4.3.	COMPARACIÓN DE LAS METODOLOGIAS AG Y FCM-AG	65
4.4.	AGRUPACIÓN CON GBCM	69
4.5.	COMPARACIÓN RESULTADOS FCM-AG Y GBCM	71
4.6.	ANÁLISIS DE TRANSFORMADORES SOSPECHOSOS	75
5.	TRABAJOS FUTUROS	81
6.	BIBLIOGRAFIA	82
7.	ANEXO – PARAMETRIZACIÓN DEL AG	89
7.1.	POBLACIÓN INICIAL Y TASA DE MUTACIÓN	89
7.2.	TIPO DE MUTACIÓN Y TASA DE <i>CROSSOVER</i>	91
7.3.	CRITERIO DE PARADA	91

LISTA DE FIGURAS

	Pág.
Figura 1. Registro histórico de pérdidas para un transformador de distribución....	20
Figura 2. Proceso para descubrir conocimiento en bases de datos	21
Figura 3. Proceso de agrupamiento jerárquico.....	28
Figura 4. Deducción distancia entre dos objetos en plano cartesiano.....	32
Figura 5. Esquema algoritmo genético básico.....	40
Figura 6. Esquema metodológico general	42
Figura 7. Valor atípico de pérdidas.....	44
Figura 8. Casos típicos filtrados de acuerdo a un análisis de pre-procesamiento.	45
Figura 9. Diagrama esquemático del AG	48
Figura 10. Población inicial del AG.....	48
Figura 11. Esquema de asignación de población inicial.....	49
Figura 12. Esquema selección por torneo	51
Figura 13. Opción uno de mutación	53
Figura 14. Opción 2 de mutación	53
Figura 15. Opción 3 de mutación	54
Figura 16. Opción 4 de mutación	54
Figura 17. Diagrama esquemático del FCM-AG.....	56
Figura 18. Esquema GBCM	57
Figura 19. Centroides obtenidos aplicando AG	60
Figura 20. Distribución de series en la clasificación AG.....	60
Figura 21. Centroides obtenidos aplicando FCM-AG	61
Figura 22. Distribución de series en la clasificación FCM-AG	62
Figura 23. Cluster obtenidos por FCM-AG	63
Figura 24. Conmutación series de tiempo	64
Figura 25. Centroides generados por las metodologías AG y FCM-AG.....	65
Figura 26. Ejemplo conceptos de agrupación	67
Figura 27. Dendograma obtenido de aplicar el GBCM.....	70
Figura 28. Transformadores analizados por clusters 8-4, 9-4, 10-4 y 11-4	76
Figura 29. Transformadores clasificados en clusters 14 y 15 GBCM.....	78
Figura 30. Transformador 1 analizado por error en la facturación.....	79
Figura 31. Transformador 2 analizado por error en la facturación.....	79
Figura 32. Desempeño AG 20 Individuos.....	89
Figura 33. Desempeño AG 52 Individuos.....	90
Figura 34. Desempeño AG 154 Individuos.....	90

LISTA DE TABLAS

	Pág.
Tabla 1. Caracterización balances iniciales	43
Tabla 2. Caracterización datos iniciales después de etapa de pre-procesamiento	46
Tabla 3. Resumen ejecución AG	61
Tabla 4. Resumen ejecución FCM-AG	62
Tabla 5. Porcentaje de pertenencia transformadores con conmutación.....	65
Tabla 6. Distribución de agrupación de las series con AG y FCM-AG	66
Tabla 7. Índices de comparación FCM-AG y AG.....	68
Tabla 8. Comparación ejecución FCM-AG y AG	68
Tabla 9. Equivalente de agrupación entre el GBCM y el FCM-AG	72
Tabla 10. Resumen de clusters obtenidos por FCM-AG y GBCM.....	73
Tabla 11. Índices de comparación FCM-AG y GBCM	74
Tabla 12. Análisis clusters 8-4, 9-4, 10-4 y 11-4	75
Tabla 13. Parametrización Tasa de Crossover y tipo de mutación	91
Tabla 14. Parametrización criterio de parada.....	92

1. PROBLEMA DE INVESTIGACIÓN

1.1. PLANTEAMIENTO DEL PROBLEMA

Las pérdidas de energía eléctrica son una problemática que se presenta tanto en las empresas de distribución como en la sociedad misma. Se pueden clasificar en función de su naturaleza así:

- 1) Pérdidas técnicas: Se producen principalmente en equipos de medida, transporte y transformación. Se asocian a fenómenos físicos naturales ya que son producidas por disipación de energía en forma de calor y son consecuencia directa de la cantidad de energía que se transporta desde los centros de generación hasta los centros de consumo.
- 2) Pérdidas no técnicas: Es la energía que se dispone por parte de las empresas para el consumo de los usuarios, pero que por factores externos al sistema eléctrico no se factura. Entre estas se encuentran los fraudes de energía, las anomalías sobre los sistemas de medida y/o errores administrativos en la facturación.

El mayor porcentaje de pérdidas no técnicas se concentra en los fraudes y las anomalías, especialmente en las redes de baja tensión. Como consecuencia, en países de América Latina y el Caribe el 80% de las pérdidas totales ocurren a nivel de distribución, principalmente por aspectos no técnicos [1].

Los fraudes de energía se definen como la acción ilegal llevada a cabo por el usuario con el fin de reducir la cantidad de energía cobrada por la empresa comercializadora, esta acción infiere una facturación errónea del consumo. Por otro lado, una anomalía se asocia también a una facturación errónea del consumo a causa de un daño sobre el sistema de medida, se diferencia del fraude en que no se presenta con dolo por parte del usuario. Aunque los fraudes y las anomalías son inevitables debido a fenómenos culturales, sociales y a situaciones de tipo estocásticas, se pueden realizar acciones para mitigar su efecto, las cuales se pueden enmarcar en el uso de herramientas tecnológicas o involucrando un esfuerzo humano [2].

La gestión de pérdidas trae innumerables beneficios, entre los que se consideran: recuperación de costos y sostenibilidad financiera de las empresas, capitalización del sector eléctrico y mejora en la capacidad de inversión, disminución en la emisión de gases de efecto invernadero y reducción en las tarifas eléctricas para el consumidor final.

Como alternativa para el control y reducción de pérdidas de energía las empresas han instalado medidores (macromedidores¹) a las salidas de los transformadores de distribución con el objetivo de realizar balances de energía² y determinar transformadores objetivo para inspección en campo. Los balances por transformador se calculan mensualmente y se analizan en la mayoría de empresas con métodos manuales poco estructurados.

Debido a la cantidad de datos que se pueda generar, a la complejidad del patrón de comportamiento histórico de las pérdidas en los transformadores y a la manualidad del análisis, se dificulta formular estrategias y tomar decisiones efectivas con esta información. Se requiere entonces una metodología basada en algoritmos computacionales que facilite el análisis de los datos para identificar las pérdidas en los sistemas de distribución.

La minería de datos es una de las técnicas más utilizadas actualmente para descubrir información sobre las crecientes y desaprovechadas bases de datos de las empresas. Algunas de las ventajas que se asocian al proceso de minería de datos son: Descubre información de manera automática sobre bases de datos, los resultados obtenidos son de fácil interpretación, contribuye a la toma de decisiones tácticas y estratégicas de una organización y supone ahorro de costos para las empresas.

1.2. JUSTIFICACIÓN

El análisis manual y poco estructurado de los datos induce baja efectividad en el direccionamiento de acciones en terreno para la identificación de pérdidas no técnicas, lo que implica un sobre costo en los planes de control y reducción de pérdidas de energía, haciendo de estos una actividad no sostenible a largo plazo para la empresa.

En países como Colombia, Chile, Costa Rica y Perú, se han implementado esquemas regulatorios que “castigan” o “recompensan” a las empresas por su gestión sobre las pérdidas [1]. Para el caso específico de Colombia, se ha propuesto reconocer un índice de referencia en el nivel de tensión 1 a las empresas distribuidoras con gestión de pérdidas más eficiente, incluyendo su costo en la tarifa que se cobra al usuario final. Si las empresas distribuidoras sobrepasan este nivel de referencia deben asumir una reducción directa sobre sus ingresos. Así se asegura el control de pérdidas en empresas que han alcanzado cierto nivel de eficiencia. Para las empresas que no han alcanzado el nivel de eficiencia propuesto,

¹ Medidores de energía instalados en los bornes secundarios de un transformador de distribución con el fin de totalizar la energía total entregada a la carga.

² Corresponde a la diferencia entre el consumo del macromedidor y el consumo total de los usuarios asociados a un transformador.

se les permite implementar planes de reducción de pérdidas cuyo costo se reconocerá basado en un cumplimiento gradual de metas [3].

Es una necesidad de las empresas de energía buscar nuevas alternativas que les permita gestionar las pérdidas no técnicas de manera eficiente, con el fin de ser competitivas en el mercado, contribuir de manera positiva frente al impacto ambiental y ser sostenibles a largo plazo.

La minería de datos surge como una necesidad de la sociedad actual de implementar técnicas y estrategias informáticas que permitan, de forma automática, el análisis de grandes cantidades de datos que a diario se almacenan en bases de datos poco exploradas. Con la minería de datos se pueden identificar patrones de comportamiento relevantes, describir tendencias y predecir comportamientos. Las técnicas basadas en *Clustering*³ han sido ampliamente utilizadas en diversos campos, entre ellos las pérdidas de energía [4].

1.3. ANTECEDENTES

La minería de datos (*data mining*) es un campo de la estadística y las ciencias de la computación que surge como una necesidad de la sociedad actual de implementar técnicas y estrategias informáticas eficientes para el análisis de datos. Los objetivos que se buscan con la minería de datos son: descubrir patrones de comportamiento relevantes, describir tendencias y predecir comportamientos. La minería de datos hace parte fundamental del proceso de descubrimiento de información (KDD, *Knowledge Discovery in Databases*), el cual consta de varias etapas e incorpora diferentes técnicas.

En un proceso típico de KDD se siguen los siguientes pasos: 1. Selección del conjunto de datos 2. Análisis de las propiedades de los datos 3. Transformación del conjunto de datos de entrada 4. Selección y aplicación de la técnica de minería de datos 5. Extracción de conocimiento y 6. Interpretación y evaluación de los resultados [4].

Las técnicas de minería de datos son algoritmos que provienen de la inteligencia artificial y la estadística y se aplican sobre un conjunto de datos para obtener unos resultados. Algunas técnicas representativas son: Redes neuronales, Regresión lineal, Árboles de decisión, Modelos estadísticos, Agrupamiento o *Clustering* y Reglas de asociación.

La técnica de agrupamiento o *Clustering* se refiere al proceso de agrupar un conjunto de objetos en clases similares. Por otro lado, un *cluster*⁴ es una colección de datos u objetos parecidos entre si y diferentes a los almacenados en otros *cluster*

³ Algoritmo basado en minería de datos para el agrupamiento de objetos por clases similares

⁴ Conjunto de elementos con clases similares, formado a partir de un algoritmo de agrupamiento

o grupos. El *clustering* ha sido abordado por varias disciplinas y se aplica a una variedad de contextos, reflejando la utilidad que tiene como uno de los pasos en el análisis experimental de los datos. Sus resultados no dependen de clases previamente definidas, ni de ejemplos de entrenamientos etiquetados a priori, por esta razón se señala que su aprendizaje es de tipo “no supervisado”⁵, porque aprende por observación en vez de por ejemplos. Las técnicas de *clustering* han sido utilizadas en múltiples aplicaciones tales como: reconocimiento de patrones, análisis de datos, procesamiento de imágenes, estudio de mercados, entre otras [5- 6].

Existe un gran número de algoritmos de *clustering*, los cuales se clasifican por categorías: 1. Métodos particionales, 2. Métodos jerárquicos, 3. Métodos basados en densidad, 4. Métodos basados en *grids* y 5. Métodos basados en modelos [7].

Los métodos particionales se caracterizan por clasificar los n datos iniciales en k *clusters*, cumpliendo las siguientes condiciones: 1. La similitud entre miembros de un mismo *cluster* es elevada (similitud *intra-cluster*). 2. La similitud entre miembros de diferentes *cluster* es baja (similitud *inter-cluster*). 3. Cada *cluster* debe contener como mínimo un elemento. 4. Cada elemento debe pertenecer únicamente a un grupo (*Hard*), de esta manera $k \leq n$. La cuarta condición se relaja para ciertas técnicas, así nacen las técnicas de *clustering* difusas (*Fuzzy Clustering*). Algunas de las técnicas más populares en los métodos particionales son el Algoritmo de *k-means* [8] y el algoritmo de *Fuzzy C-means* [9].

El *Fuzzy C-means* (FCM) es un algoritmo de agrupamiento difuso, donde cada elemento tiene un grado de pertenencia difuso a los grupos, representado por una función de pertenencia que toma valores entre cero y uno, donde cero es similitud nula y uno es similitud total. La suma de los valores de pertenencia de un elemento sobre todos los *clusters* es igual a la unidad. En forma general, el algoritmo FCM define una función objetivo y opera de forma iterativa minimizando esta función hasta alcanzar una partición difusa óptima. Al minimizar esta función objetivo se maximiza la similitud entre miembros de un mismo *cluster*, mientras que se minimiza la similitud entre los diferentes grupos. FCM tiene la bondad de ser simple en su implementación, se puede usar en diversos conjuntos de datos, requiere poca memoria para su ejecución y converge de manera rápida, casi lineal. Sin embargo, es altamente dependiente de los valores de inicio que se especifiquen, por lo cual no garantiza una convergencia a óptimos globales [10].

Como alternativa de solución al problema de convergencia local del FCM se han implementado diferentes estrategias para definir mejores condiciones iniciales, las cuales aseguran una convergencia cercana al óptimo global. Entre éstas se destacan estrategias de *clustering* basadas en algoritmos evolutivos [11], [12] y [13].

⁵ Método de aprendizaje automático para el cual no se tiene un conocimiento previo de la situación

Los algoritmos genéticos (AG) se enmarcan dentro de los algoritmos evolutivos, actúan sobre una población de soluciones candidatas haciéndola evolucionar en cada iteración mediante acciones aleatorias propias de la naturaleza (mutación, recombinación). Cada nueva iteración representa una generación, en la cual se evalúa cada individuo mediante una medida de aptitud que simula la función objetivo en el problema de optimización a resolver. Los individuos más aptos son seleccionados y se someten nuevamente a un proceso de evolución para formar una nueva generación. El proceso termina cuando se ha alcanzado un número pre-establecido de iteraciones o cuando se ha alcanzado un nivel de aptitud satisfactorio. Los AG se hicieron populares a partir de la publicación de John Henry Holland en los años setenta [14].

El primer trabajo en utilizar una asociación directa de *clustering* con AG fue realizado por Raghavan *et. Al* [15]. El objetivo principal de esta técnica híbrida es determinar un punto representante de cada *cluster* y asignar cada objeto a su *cluster* más cercano. Muchos estudios han mostrado que este enfoque es más robusto en la convergencia a la partición óptima que los algoritmos de partición clásicos [16], [17], [18] y [19].

Algunos trabajos se han enfocado en desarrollar algoritmos de *clustering* híbridos, combinando técnicas clásicas de *clustering* con AG, en los cuales cada solución candidata del AG se usa como punto de partida de un método particional. La calidad de la solución encontrada por el método particional se usa como función *fitness*⁶ de la solución candidata del AG [20].

Por otra parte se encuentran los métodos de agrupación jerárquica (*Hierarchical Clustering*), los cuales son algoritmos de agrupación que se ejecutan por niveles, en cada nivel se hace división o fusión de *clusters* hasta alcanzar un número de agrupaciones deseadas. Los resultados obtenidos pueden ser graficados en forma de árbol, al cual se denomina dendograma. En el dendograma se observa todo el proceso intermedio de división o fusión de *clusters*, hasta llegar a los resultados finales. De acuerdo si el proceso se ejecuta mediante división o fusión, los métodos jerárquicos se pueden clasificar en divisivos o aglomerativos, respectivamente [21].

Al igual que los métodos particionales, los métodos de agrupamiento jerárquico se basan en una medida de similitud⁷ para realizar la agrupación, diferenciándose en que los métodos particionales realizan agrupamiento de los datos en un solo nivel (una etapa). Sin embargo, para el algoritmo jerárquico es posible aplicar en cada etapa un esquema de agrupación particional para realizar la agrupación [22].

⁶ Función utilizada en los algoritmos genéticos para mejorar la selectividad entre los diferentes individuos candidatos a solución, con el fin de que el algoritmo diferencie los individuos de mejor calidad de los de peor calidad.

⁷ Criterio definido para medir la semejanza entre dos elementos.

El *Bisecting K-Means* (B-KMA) es un algoritmo jerárquico divisivo que utiliza el método de *Clustering* particional *k-Means* para obtener la mejor agrupación de *clusters* en cada etapa [23-24]. El algoritmo, por ser de naturaleza divisiva, inicia con todos los objetos agrupados en un mismo *cluster*, hasta alcanzar el nivel de división propuesto. El B-KMA divide recursivamente un *cluster* en dos *sub-clusters*, partiendo de un conjunto de datos inicial formado por todos los elementos. Los *sub-clusters* se obtienen de aplicar en cada etapa al *cluster* padre el algoritmo de *k-Means* con $k=2$. Al realizar divisiones recursivas se pueden obtener al final tantos *clusters* como números de datos iniciales, formando en el proceso un árbol jerárquico en forma de dendograma.

Finamente, los trabajos de investigación consultados sobre detección de pérdidas no técnicas de energía a través de minería de datos, se centran principalmente en el análisis del patrón de comportamiento del consumo histórico de los clientes (perfil de carga). A partir del perfil de carga del cliente se detectan desviaciones significativas en el comportamiento histórico del consumo, que se asocian a la presencia de pérdidas no técnicas. Los métodos de análisis aplicados en estos trabajos son de tipo supervisados y no supervisados.

Los métodos supervisados son los más frecuentes, pero para ser aplicados requieren de un conjunto de datos de entrenamiento con una etiqueta en la que se referencia si el cliente ha presentado o no problemas de pérdidas no técnicas. Esta etiqueta es difícil de conseguir en las bases de datos reales, ya que no se tiene una cultura en las empresas que permita estructurar los datos de esta manera. Por su parte, los métodos no supervisados no requieren de un conjunto de entrenamiento para su aplicación, asignan una clasificación al cliente a partir de una inferencia estadística [21].

Algunos trabajos destacados en los que se han aplicado diferentes métodos para la detección de pérdidas no técnicas de energía son:

- En [25] se presenta una técnica para el análisis de pérdidas no técnicas (PNT) usando *Support Vector Machine* (SVM). La técnica busca preseleccionar clientes sospechosos de fraude para ser inspeccionados en campo a partir de la predicción de patrones de consumo anormales. En la investigación se utiliza la información del perfil de carga del usuario para determinar un comportamiento anormal relacionándolo con un marco de análisis de PNT propuesto, entre los que se incluyen: la manipulación al medidor y errores en la medición, entre otras. Las pruebas aplicadas muestran que usar SVM en la detección de fraudes puede ser muy efectivo dado que éste tiene una división no lineal de hiper-superficies que le da alta discriminación y tiene una alta capacidad de generalización para la clasificación de datos ocultos. La tasa de efectividad alcanzada con el sistema de detección propuesto es del 50%.

- En [26] se presenta el *Support Vector Clustering* (SVC) como una estrategia de clasificación de patrones de carga de un conjunto de datos. Los resultados ofrecidos por el SVC muestran la capacidad del método para aislar patrones de carga no comunes y formar grupos de patrones de carga consistentes. La formación de agrupaciones explota el significado conceptual de los vectores de soporte acotadas como valores atípicos y utiliza esta información para limitar la búsqueda de los puntos de datos que se asignarán a la misma agrupación. Los resultados obtenidos en un gran conjunto de datos de clientes de electricidad reales muestran que el algoritmo propuesto es competitivo con las técnicas de agrupación (*clustering*).
- En [27] se presenta un enfoque para la detección de Pérdidas no técnicas (PNT), mediante la combinación de tareas de minería de datos, algoritmos de agrupamiento (*Clustering*) y técnicas de clasificación. Para ello se involucran tres módulos principales: Agrupación: El comportamiento de carga del cliente individual se agrupa en conjuntos de comportamientos iguales. Detección: Se utiliza el aprendizaje de reglas para especificar si los datos son PNT o No PNT. Clasificación: Se genera un modelo de clasificación con el fin de predecir el comportamiento futuro del cliente que sigue ciertos patrones.
- En [28] se aplica el método *Rough Sets* para predecir perfiles de consumo de usuarios fraudulentos, usando un conjunto de reglas de clasificación. En el artículo, *Rough sets* se utilizó para eliminar atributos condicionales y el MDA (*minimal decision algorithm*) se utilizó para eliminar clases insignificantes de cada atributo condicional. La minimización de la base de datos enfocada al comportamiento de los usuarios, permite obtener unas reglas de clasificación en el sistema para predecir perfiles de consumo de usuarios tendientes al fraude. Se concluye en el artículo que *Rough Sets* es una poderosa herramienta para la detección de fraudes, especialmente cuando no existe ningún conocimiento previo del sistema pero si de la base de datos. Se concluye también que uno de los principales problemas de los métodos de clasificación utilizados en la detección de fraudes es su alta dependencia a la calidad de los datos analizados.
- En [29] se utilizan dos métodos basados en minería de datos para la búsqueda de clientes con fraude, en el primer método realiza la búsqueda basado en un proceso de agrupación con árboles de decisión y se aplica a clientes con patrones de consumo similares a clientes infractores detectados previamente. El segundo método detecta los clientes con caídas considerables en el consumo en relación con periodos anteriores. Los métodos se prueban sobre sistemas reales alcanzando una efectividad de hasta el 20%.
- En [30] se construye un sistema de reconocimiento de patrones con el fin de identificar consumidores de energía con comportamientos sospechosos de

fraude. El parámetro de reconocimiento de patrones se describe en detalle y se plantea el algoritmo como un problema de optimización.

- En [31] se da una mirada integral a las pérdidas no técnicas, los perfiles de carga y las técnicas de minería de datos utilizados para minimizar las pérdidas de energía. El objetivo principal del artículo es predecir y detectar pérdidas de energía no técnicas en sectores del SDL usando minería de datos sobre los perfiles de carga de los usuarios. El artículo se centra en el análisis del comportamiento histórico de consumo de los clientes.
- En [32] se compara el rendimiento de dos metodologías de clasificación supervisadas (Naïve Bayes y árboles de decisión), mediante las cuales se extrae el patrón de comportamiento histórico del consumo de los clientes para luego ser usado en la detección de usuarios con problemas de pérdidas no técnicas.
- En [33] se presenta una metodología híbrida supervisada para el análisis de pérdidas no técnicas utilizando Algoritmos Genéticos y Máquinas de soporte vectorial. En este trabajo se seleccionan clientes sospechosos para inspección en campo, a partir del comportamiento anormal del consumo histórico del cliente.
- En [34-35] se presenta una metodología no supervisada para identificar usuarios infractores. La metodología se desarrolla en dos pasos. En el primero se utiliza el algoritmo *Fuzzy C-means* (FCM) para agrupar a los clientes en *clusters* con perfiles de consumo similares. Producto de la clasificación con el FCM se obtuvo una matriz de pertenencia difusa⁸ y la distancia de cada usuario a cada uno de los centros de los grupos formados⁹. En el paso dos se ordenan los usuarios a partir de la distancia calculada y se asigna un índice de puntuación unitario. Se asume que los usuarios con patrones de comportamiento irregulares tienen la más alta puntuación.
- En [2] se presenta un enfoque para el análisis de pérdidas no técnicas usando “*Extreme Learnin Machine*” (ELM), mediante el cual se extraen patrones de comportamiento del perfil de carga histórico de los clientes. Los perfiles de carga son clasificados con base a los patrones observados y se generan categorías que son usadas para revelar si alguno de los comportamientos significativos obtenidos se debe a irregularidades en el consumo.
- En [36] se utilizan dos técnicas de aprendizaje automático supervisado (*Back-Propagation Neural Network* (BPNN) y *Online secuencial Machine Learning Extreme* (OS-ELM)) para la detección de pérdidas no técnicas, usando como datos de entrada el consumo histórico de los clientes. Como aporte del mismo,

⁸ Matriz en la que se relaciona el porcentaje de pertenencia de cada cliente a cada uno de los grupos formados

⁹ Distancia Euclidiana a los centroides de cada *cluster*

se comparan los resultados obtenidos con los resultados aplicando Máquinas de Soporte Vectorial.

La mayoría de las referencias revisadas se basan en métodos supervisados y el insumo principal de análisis son los consumos históricos de los clientes. No se encontró en las referencias un trabajo cuyo insumo sea el consumo de los macromedidores, o una combinación entre el consumo de los macromedidores, el consumo total de los usuarios y el registro histórico de las pérdidas asociadas al transformador.

1.4. OBJETIVOS

1.4.1. Objetivo general

Identificar pérdidas de energía en transformadores de distribución a partir de una metodología híbrida no supervisada basada en algoritmos genéticos (AG) y *Fuzzy C-Means* (FCM).

1.4.2. Objetivos específicos

1. Definir la función de distancia en el FCM para medir la semejanza entre patrones de datos con temporalidad.
2. Implementar un método particional difuso no supervisado para la agrupación de series de tiempo.
3. Implementar una estrategia híbrida particional, basada en algoritmos evolutivos y en un algoritmo particional difuso.
4. Implementar una estrategia jerárquica divisiva difusa para la agrupación de series de tiempo.
5. Proponer un índice adecuado para evaluar la validez de los *clusters* encontrados a partir de la estrategia híbrida FCM-AG.
6. Plantear una parametrización adecuada del AG, para que, operando en conjunto con el FCM garantice una buena solución del problema a abordar.
7. Encontrar comportamientos de las series agrupadas que permitan inferir transformadores con mayor tendencia al fraude.
8. Plantear y evaluar los resultados de los procedimientos de agrupación propuestos con el fin de validar las soluciones obtenidas.

1.5. ESTRUCTURA DEL DOCUMENTO

El presente trabajo se estructura en seis capítulos. En el capítulo 2 se presenta el marco teórico del problema de investigación a abordar, se referencian algunas definiciones importantes y se formulan algunos métodos de agrupación por *clustering* que servirán más adelante para el desarrollo de la metodología propuesta. En el capítulo 3 se expone la metodología propuesta, iniciando por la adquisición de los datos, hasta llegar a la etapa de procesamiento de los mismos. En el capítulo 4 se muestran los resultados obtenidos de aplicar la metodología propuesta sobre una base de datos real y se hace una interpretación de los resultados obtenidos. En el capítulo 5 se proponen diferentes líneas de investigación derivadas del presente trabajo. Finalmente, en el capítulo 6 se incluyen las fuentes de información consultadas.

2. MARCO TEÓRICO

2.1. LA MACROMEDICIÓN Y EL BALANCE DE ENERGÍA

La macromedición hace referencia a la medida de la energía total que circula por los bornes de salida de los transformadores de distribución hacia la carga. Por razones económicas, en Colombia la macromedición es instalada principalmente en los bornes secundarios de los transformadores (Nivel de tensión 1). De esta manera la energía registrada por el macromedidor corresponde a:

1. Consumo de energía de los usuarios aguas abajo del transformador (*C.Usuarios*).
2. Consumo del alumbrado público y otras cargas de cobro por aforo (*C.AP ~ O.C.*)
3. Pérdidas técnicas (Redes de nivel 1, medidores y acometidas).
4. Pérdidas no técnicas (Fraudes y anomalías).

El balance de energía en un transformador de distribución con macromedidor corresponde a:

$$C.Macromedidor = C.Usuarios + C.AP \sim O.C + Perd.Tecnicas + Perd.NoTecnicas \quad (1)$$

Asumiendo:

$$Perd.Totales = Perd.Tecnicas + Perd.NoTecnicas$$

Se concluye que las pérdidas totales en el transformador corresponde a:

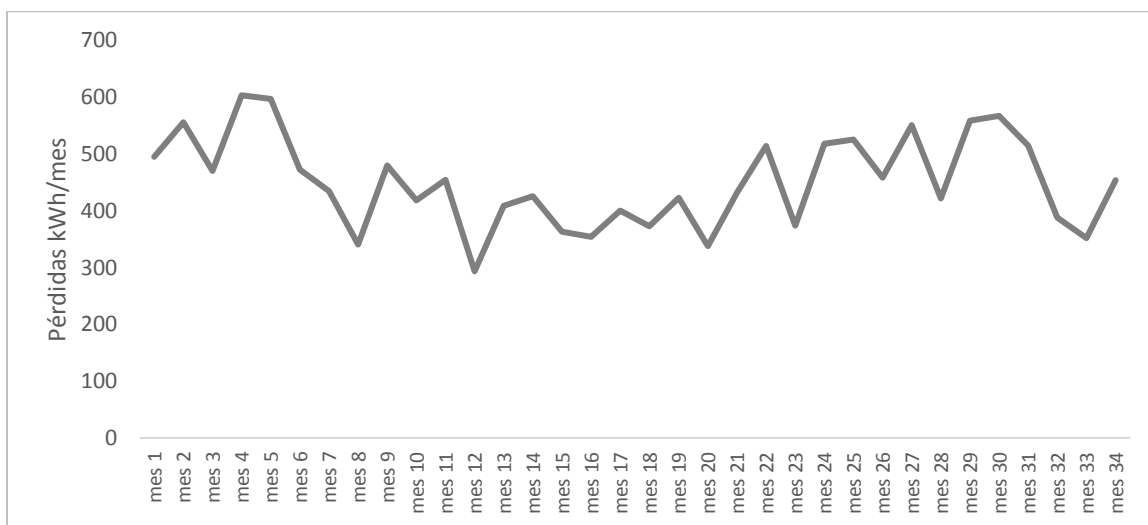
$$Perd.Totales = C.Macromedidor - C.Usuarios - C.AP + OC \quad (2)$$

A partir de la ecuación (2) se calculan las pérdidas totales en los transformadores de distribución con macromedición instalada. Este cálculo es importante para la gestión de pérdidas, especialmente cuando la cobertura de macromedición tiende a cubrir todo el parque de transformadores del sistema de distribución. Aunque es alta la inversión para implementar macromedición fija sobre todos los transformadores del sistema de distribución, se presenta como una necesidad para los operadores de red en Colombia ya que se debe cumplir el artículo 146 de la ley

142 de 1994 [37]: “...las empresas tendrán un plazo a partir de la vigencia de la presente Ley para elevar los niveles de macro y micromedición a un 95% del total de los usuarios...”. También, en la resolución CREG 172 de 2011, se establece la instalación de macromedición en transformadores de distribución como una de las actividades a desarrollar en la implementación de un plan de reducción de pérdidas no técnicas [3].

La macromedición fija en los transformadores permite construir un registro histórico del comportamiento de las pérdidas, con el cual se diseñan estrategias de control y reducción. En la Figura 1 se muestran las pérdidas de un transformador (balance de energía) durante un periodo de 34 meses.

Figura 1. Registro histórico de pérdidas para un transformador de distribución



Ya que los sistemas de distribución son dinámicos en el tiempo y se afectan por eventos estocásticos, se deben garantizar algunas condiciones del balance de energía para la toma de decisiones a partir de los perfiles de pérdidas obtenidos. Estas son:

- Sincronizar y actualizar permanentemente el inventario de usuarios asociados al transformador, de tal forma que permita reflejar en el sistema comercial de la empresa la realidad de lo que sucede en terreno.
- Actualizar recurrentemente en los transformadores el aforo¹⁰ de alumbrado público y otras cargas de este tipo, de tal forma que se refleje la realidad del consumo con la carga aforada.

¹⁰ Facturación de consumo de energía basado en la potencia de consumo y el tiempo de uso

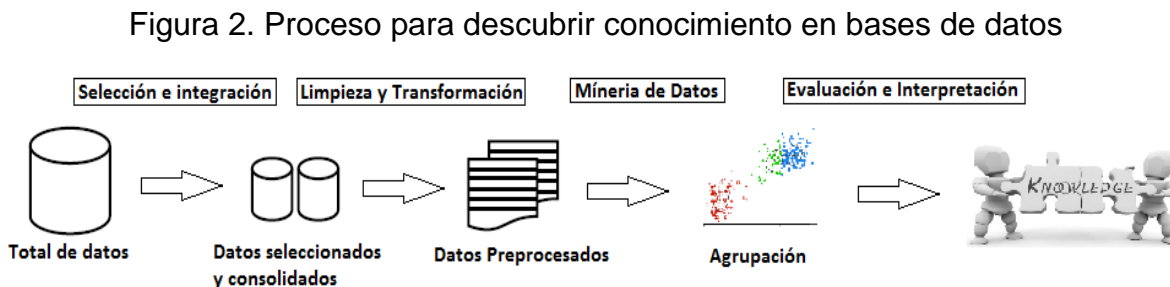
- Sincronizar la toma de lectura del macromedidor y los usuarios asociados a un mismo transformador, garantizando que transcurra el menor tiempo posible para ello.
- Garantizar el correcto funcionamiento del sistema de medida, tanto del macromedidor como de los usuarios.

Puesto que a todos los transformadores del sistema se les realiza mensualmente un balance de energía, el volumen de información que se genera es muy alto y con tendencia creciente, por lo cual se dificulta el análisis de la misma a través de métodos manuales. Por lo anterior, se visualiza la necesidad de explorar técnicas que permitan analizar la información de manera automática y en modo *offline*¹¹, permitiendo concentrarse en la toma de decisiones más que en el procesamiento de los datos.

2.2. EXTRACCIÓN DE CONOCIMIENTO EN BASES DE DATOS (KDD)

En los últimos años el mercado de la tecnología ha permitido el progreso constante del mundo de la informática, permitiendo que hoy se tengan desarrollos tanto de hardware como de software capaces de almacenar y procesar grandes cantidades de datos. Este crecimiento continuo de los datos excede la capacidad de los seres humanos de analizarlos sin el uso de ayudas externas. La creciente brecha entre los datos y la información, útil para la toma de decisiones, requiere el desarrollo de algoritmos que permitan convertir las bases de datos en conocimiento tangible [38].

La Extracción de conocimiento está principalmente relacionada con el proceso de descubrimiento conocido como *Knowledge Discovery in Databases* (KDD), que se refiere al proceso no-trivial de descubrir conocimiento e información potencialmente útil dentro de los datos contenidos en algún repositorio de información [7]. El término proceso involucra una serie de etapas iterativas, las cuales van desde la preparación de los datos, la extracción de patrones (minería de datos) y la evaluación e interpretación de los resultados. El proceso de descubrimiento de conocimiento se muestra en la Figura 2 [23].



¹¹ Fuera de línea. No se realiza en tiempo real.

En este proceso los datos son la materia prima, a los cuales se les asigna un atributo¹² para que posteriormente se conviertan en información. Cuando un modelo se ajusta a los datos, permitiendo obtener una interpretación de estos y generando valor agregado, se obtiene un resultado llamado conocimiento.

2.2.1. La minería de datos

La minería de datos es parte fundamental del proceso KDD y se refiere a la etapa en la cual se modelan los datos a partir de la aplicación de métodos inteligentes, con el objetivo de extraer patrones de comportamiento, describir tendencias, descubrir agrupaciones naturales, extraer registros poco usuales o anómalos e identificar dependencias entre los datos en un determinado contexto.

Asociado a la etapa de minería de datos se involucran técnicas que provienen de múltiples disciplinas, las más representativas son:

- Redes Neuronales
- Algoritmos genéticos
- Regresión lineal
- Árboles de decisión
- Modelos estadísticos
- *Clustering*
- Procesos de asociación

Finalmente, se subraya que la minería de datos es considerada uno de los desarrollos interdisciplinarios más prominente en la tecnología de la información [40].

2.2.2. Algoritmos de agrupamiento en tareas de *Clustering*

Los algoritmos de agrupamiento o *Clustering* hacen parte del campo de aprendizaje automático (*Machine Learning*) cuyo objetivo es desarrollar técnicas que permitan “aprender” a las computadoras. Los algoritmos de *Clustering* se enmarcan dentro del aprendizaje “No supervisado” y se constituyen como el proceso de agrupación de elementos o variables en clases similares llamadas *cluster*, de tal forma que los *clusters* capturen la estructura natural de los datos, buscando la mayor homogeneidad al interior de cada uno de los grupos formados y alta diferencia entre ellos.

El análisis de agrupamiento no tiene bases sobre las que deducir inferencias estadísticas para una población a partir de una muestra, es un método basado en

¹² Especificación que define una propiedad de un objeto

criterios geométricos y se utiliza fundamentalmente como una técnica exploratoria, descriptiva pero no explicativa. Las soluciones no son únicas, en la medida en que la pertenencia al *cluster* para cualquier número de soluciones depende de muchos elementos del procedimiento elegido. Por otra parte, la solución *cluster* depende totalmente de las variables utilizadas, la adición o destrucción de variables relevantes puede tener un impacto substancial [41].

El *clustering* es actualmente una disciplina en desarrollo con significativos estudios en diversos campos de aplicación entre los cuales se destacan la minería de datos, estadística, *Machine Learning*, *marketing*, etc. Dentro de los desafíos a sortear en los diferentes campos de aplicación del *clustering* se encuentran [42]:

- Escalabilidad: Las técnicas de *Clustering* trabajan bien sobre bases de pequeña escala. Cuando la base de datos es significativa en relación al número de objetos los resultados obtenidos tienden a ser parciales, sacrificando una solución de tipo global.
- Tipo de atributos: Las técnicas de *Clustering* están diseñadas en su mayoría para ser aplicadas sobre atributos de tipo numérico, sin embargo de acuerdo al área de aplicación se requiere que trabajen con otro tipo de atributos (binarios, nominales, etc).
- Forma, densidad y tamaño del Clúster: Los algoritmos *Clustering* realizan agrupamiento de datos basados en una medida de similitud, en su mayoría utilizan medidas de distancia. Estas medidas tienden a formar *clusters* de forma esférica y de tamaño y densidad similar. En aplicaciones reales los datos no necesariamente tienden a formar grupos simétricos, por lo cual se requiere el desarrollo de algoritmos que utilicen diferentes criterios para la agrupación.
- Parámetros de entrada: La mayoría de los algoritmos de *Clustering* utilizan como parámetro inicial el número de *clusters* k en el que se clasifican los datos y/o los centroides¹³ V_j de cada grupo. Estas técnicas por ser de explotación local son muy sensibles a los parámetros de entrada, los cuales para aplicaciones reales tienden a ser desconocidos.
- Datos atípicos o erróneos: El *Clustering* tienden a ser sensibles a este tipo de datos (*Outliers*¹⁴, datos vacíos, etc), lo que sacrifica la calidad en las agrupaciones obtenidas.
- Dimensión de los datos: Los algoritmos de *Clustering* tienen un buen desempeño en dos o tres dimensiones, sin embargo las bases de datos reales en general

¹³ Punto equidistante a todos los objetos pertenecientes a un mismo *cluster*

¹⁴ Observación distante del resto de los datos. Valor atípico.

tienen dimensiones mayores, lo que se convierte en un desafío por la diversidad entre los datos.

- Interpretabilidad y usabilidad: Los resultados obtenidos se espera sean de fácil interpretación y útiles para la toma de decisiones.

Existen varios algoritmos de *Clustering*, la aplicación de uno u otro requiere de un análisis previo de la base de datos, del recurso computacional disponible y del tipo de resultados que se buscan. Estos algoritmos se pueden clasificar en las siguientes categorías generales [42]:

- Métodos particionales: Los métodos particionales clasifican los datos en k *clusters*, basados en una medida de similitud. Cada *cluster* tiene al menos un dato (no pueden existir grupos vacíos), y dependiendo si el método particional es duro (*Hard*) o difuso (*Fuzzy*), cada dato se asocia a un único *cluster* o a varios con un grado de pertenencia asignado. Generalmente estos métodos son iterativos y de explotación local¹⁵.
- Métodos jerárquicos: Estos métodos realizan una agrupación jerárquica de los datos por etapas, formando un dendograma. Dependiendo de la clasificación inicial y final, se pueden dividir en:
 - ✓ Aglomerativo: Al inicio cada dato forma un *cluster* independiente, y durante cada etapa jerárquica se van aglomerando *clusters* con características similares hasta cumplir un criterio de parada pre-establecido. Al final se puede llegar a que todos los datos pertenecen a un único clúster ($k = 1$).
 - ✓ Divisivo: Al inicio todos los datos pertenecen a un único *cluster*, en cada etapa jerárquica se van dividiendo los *clusters* en grupos, hasta alcanzar un criterio de parada pre-establecido. Al final este método puede llegar a que cada dato forma un *cluster* independiente.

Una notable desventaja de los métodos jerárquicos es que cuando se agota una etapa, está no puede deshacerse, sin embargo, se convierte también en una virtud ya que lo hace computacionalmente eficiente.

- Métodos basados en densidad: Son algoritmos de *clustering* basados en la noción de densidad de los datos como medida para la agrupación. Aquellas regiones con gran densidad de datos pueden agruparse en un *cluster*, hasta que su densidad exceda un determinado umbral. Las regiones de baja densidad de datos se pueden establecer como ruido.

¹⁵ Exploran un espacio de solución local partiendo de unos parámetros de entrada predefinidos

- Métodos basados en *Grids*: Para estos algoritmos se divide el espacio de datos en un número finito de cuadrículas (*grids*), produciendo celdas hiper-cúbicas. La metodología consiste en establecer semejanzas entre celdas para determinar *clusters*. El desempeño de estos métodos es independiente del número de datos a agrupar, sin embargo presentan alta dependencia a la dimensión de la cuadrícula [7].
- Métodos basados en modelos: Proponen un modelo para diferentes subconjuntos en los datos y luego miden la semejanza de los mismos para poder establecer los grupos existentes.

Las líneas de investigación actuales en relación al *Clustering* se centran principalmente en el estudio de la efectividad de los métodos para agrupar formas y tipos de datos complejos, desarrollo y adaptación de técnicas para *Clustering* de alta dimensión, escalabilidad de los métodos y, finalmente, métodos dirigidos a gestionar datos de naturaleza mixta en grandes bases de datos.

2.2.2.1. Algoritmo de *Clustering Fuzzy C-Means*

El *Fuzzy C-Means* (FCM, por sus siglas en ingles), es uno de los algoritmos más utilizados en el agrupamiento difuso. Fue propuesto en 1984 por Bezdek *et al.* [9] y desde entonces ha sido ampliamente citado en diversas publicaciones [43 - 44].

El FCM puede ser aplicado a datos de tipo cuantitativos (numéricos), cualitativos (categóricos) o una combinación de ambos. Para este trabajo se consideran datos cuantitativos.

Regularmente, el conjunto de datos son observaciones de fenómenos físicos, donde cada observación (objeto) consiste en una medida de n variables (atributos) almacenadas en un vector fila. Para un conjunto de N objetos, el conjunto de datos se representa por una matriz \mathbf{X} de dimensiones $N \times n$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} & \dots & x_{1n} \\ x_{21} & x_{22} & \dots & x_{2n} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ x_{N1} & x_{N2} & \dots & x_{Nn} \end{bmatrix} \quad \text{Donde: } \mathbf{X} = \{\mathbf{X}_j \mid j = 1, 2, \dots, n\}, \mathbf{X}_j \in \mathbf{R}^n \quad (3)$$

El sentido entre filas y columnas en la matriz de objetos, depende del contexto a analizar. Por ejemplo, para el caso de una empresa de distribución las filas podrían representar elementos del sistema eléctrico y las columnas causas de falla.

El FCM permite que los objetos almacenados en la matriz \mathbf{X} se agrupen en k *clusters* difusos, donde cada objeto X_j puede pertenecer a diferentes *clusters* de manera simultánea, relacionados mediante una matriz de pertenencia \mathbf{U} . Así las cosas, los objetos no se obligan a pertenecer a un solo *cluster*, sino que se les asigna un grado de pertenencia entre $[0,1]$ a cada uno de ellos, indicando su pertenencia parcial. Si el grado de pertenencia de un objeto X_j con un *cluster* es 1, entonces la pertenencia de este mismo objeto a los demás *clusters* es 0. La estructura de la matriz de pertenencia difusa $\mathbf{U}=[u_{ij}]$ se muestra a continuación:

$$\mathbf{U} = \begin{bmatrix} u_{11} & u_{12} & \dots & u_{1k} \\ u_{21} & u_{22} & \dots & u_{2k} \\ \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot \\ u_{N1} & u_{N2} & \dots & u_{Nk} \end{bmatrix} \quad (4)$$

Para esta matriz difusa se debe cumplir que:

$$\begin{aligned} u_{ij} &\in [0,1], \quad 1 \leq i \leq N, \quad 1 \leq j \leq k \\ \sum_{j=1}^k u_{ij} &= 1, \quad 1 \leq i \leq N \\ 0 < \sum_{i=1}^N u_{ij} &< N, \quad 1 \leq j \leq k \end{aligned}$$

El FCM puede ser visto como una generalización del método *k-Means*, propuesto por James MacQueen en 1967 [8]. Minimiza la siguiente función objetivo, con respecto a una función de pertenencia difusa u_{ij} y los centroides del *cluster* V_j :

$$J(\mathbf{X}; \mathbf{U}, \mathbf{V}) = \sum_{i=1}^N \sum_{j=1}^k (u_{ij})^m d^2(X_i, V_j) \quad (5)$$

El término $d^2(X_i, V_j)$ hace referencia a la medida de similitud, en términos de distancia, entre el objeto X_i y los centros de los *clusters* de agrupación V_j , denominados centroides. El peso asociado a la medida de similitud, $(u_{ij})^m$, es la m -ésima potencia del grado de pertenencia del i -ésimo dato al grupo j . Cuando m tiende a 0, la agrupación se acerca a un método particional duro. Si m se hace muy

grande, todos los objetos tienden a tener igual valor de pertenencia u_{ij} a cada uno de los k clusters formados. Cada valor de m indica un algoritmo FCM particular [9].

En [45] se realiza un análisis sobre el efecto que tiene m en la validez de las particiones obtenidas por FCM, deduciendo que el mejor valor para m se obtiene en el rango $[1.5, 2.5]$, siendo el valor medio $m = 2$ el que registra mejores resultados.

Estadísticamente la función objetivo del FCM se interpreta como una medida de la varianza total de X_i respecto a V_j . El minimizar esta función representa la solución de un problema de optimización no lineal, que se puede resolver usando una variedad de métodos (simulado recocido, algoritmos genéticos, etc), sin embargo el FCM es uno de los más utilizados [46].

El algoritmo FCM se ejecuta al seguir el siguiente proceso iterativo [10]:

Paso 1. Inicializar de forma aleatoria las funciones de pertenencia u_{ij} , tal que se cumpla:

$$\sum_{c=1}^k u_{ij} = 1$$

Paso 2. Calcular los centroides difusos V_j para $j=1,2,...,k$ usando la siguiente ecuación:

$$V_j = \frac{\sum_{i=1}^N (u_{ij})^m x_i}{\sum_{i=1}^N (u_{ij})^m} \quad (6)$$

Paso 3. Calcular la matriz de pertenencia difusa $U = [u_{ij}]$:

$$u_{ij} = \frac{\left(\frac{1}{d^2(x_i, v_j)} \right)^{\frac{1}{m-1}}}{\left(\sum_{j=1}^k \frac{1}{d^2(x_i, v_j)} \right)^{\frac{1}{m-1}}} \quad (7)$$

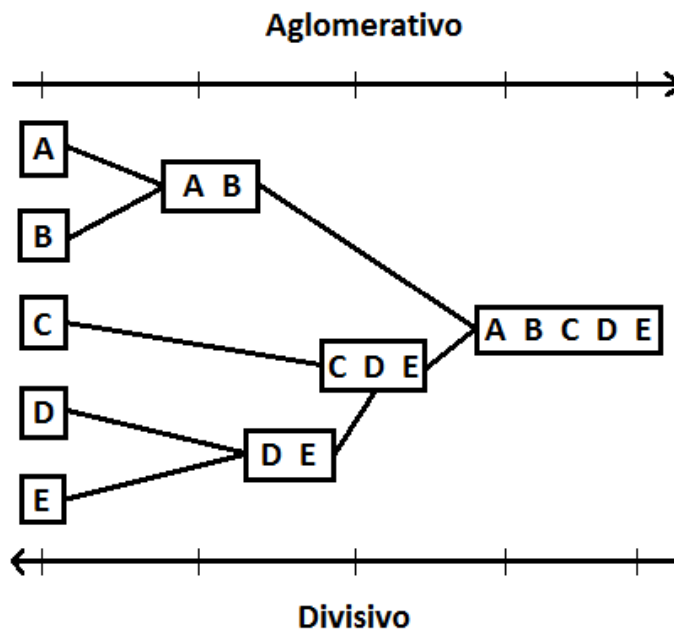
Paso 4. Repetir los pasos 2 y 3 hasta que el valor de $J(\mathbf{X}; \mathbf{U}, \mathbf{V})$ no presente un incremento considerable.

El FCM siempre converge a un óptimo local mínimo de $J(\mathbf{X};\mathbf{U},\mathbf{V})$ a partir de unas condiciones iniciales establecidas, por lo cual diferentes elecciones iniciales podrían llevar a diferentes mínimos locales.

2.2.2.2. El método jerárquico divisivo *Bisecting k-Means*

En la Figura 3 se observa un diagrama donde se ilustra el proceso de agrupamiento jerárquico.

Figura 3. Proceso de agrupamiento jerárquico



Siguiendo la Figura 3 de izquierda a derecha se observa la lógica de la agrupación jerárquica aglomerativa. Inicia con tantos *clusters* como individuos se tengan, en cada etapa jerárquica se van formando grupos basados en un criterio de similitud, hasta que al final del proceso todos los individuos se agrupen en un solo *cluster*. La lógica de la agrupación jerárquica divisiva es inversa a la del aglomerativo y se detalla siguiendo la Figura 3 de derecha a izquierda. Comienza con un conglomerado que agrupa todos los individuos, a partir de sucesivas divisiones se van formando, en cada etapa jerárquica, *clusters* cada vez más pequeños. Al final de proceso se tienen tantos *clusters* como individuos.

Como ya se mencionó anteriormente, el *Bisecting K-Means* (B-KMA) es un algoritmo jerárquico divisivo que utiliza el método de *Clustering* particional *k-Means* para

obtener la mejor agrupación en cada etapa. El algoritmo básico es como se muestra a continuación [47].

Algoritmo básico <i>Bisecting k-Means</i> para encontrar <i>k clusters</i>
<ol style="list-style-type: none">1. Inicializar la lista de <i>clusters</i>. Un grupo con todos los objetos.2. repetir3. Remover un <i>cluster</i> de la lista de <i>clusters</i>4. para $i=1$ hasta <i>numero_intentos</i> hacer5. Biseccionar el <i>cluster</i> seleccionado aplicando <i>k-Means</i>6. fin para7. Seleccionar los dos <i>clusters</i> de la bisección con el total de SSE^{16} más bajo8. Adicionar estos dos <i>clusters</i> a la lista de <i>clusters</i>.9. hasta La lista de <i>clusters</i> contenga <i>K</i> grupos.

El B-KMA es una extensión del algoritmo *K-Means*, pero visto de manera jerárquica. Inicia dividiendo el total de objetos en dos, usando para ello el *K-Means*. Luego selecciona uno de los dos *clusters* obtenidos para hacer una nueva división. De forma iterativa repite el proceso formando de a dos *clusters* en cada etapa o nivel jerárquico, hasta obtener un número de *clusters* previamente definido.

Existen varias alternativas para seleccionar el *cluster* que pasa a la etapa de bisección, una de estas es seleccionar el de mayor tamaño (el de más objetos), otra alternativa es utilizar un criterio basado en medidas de similitud y otra es utilizar una combinación de las dos. En [22] se realizaron múltiples corridas utilizando estos criterios y encontraron que las diferencias no eran sustanciales entre el uso de uno y otro.

2.2.3. Normalización de los datos para un análisis de *Clustering*

La normalización de los datos se puede asumir como una tarea de pre-procesamiento para el *clustering*. Es importante porque permite escalar o controlar la variabilidad del conjunto de datos. Aplicar una medida de distancia a datos con atributos de diferente magnitud, diferentes unidades o con un rango de variación muy alto, implica un riesgo en la formación de los *clusters*, ya que la distancia entre objetos puede ser más significativa para algunos casos particulares que para otros. En este sentido se recomienda no emplear medidas de distancia sin tener una normalización previa [48].

¹⁶ *Sum of the Squared Error (SSE)*, método de validación no supervisado de *clusters* en espacios euclidianos, que computa la distancia de todos los puntos a su centroide más cercano

El objeto de la normalización es llevar a una misma base de magnitud y/o dimensiones los atributos asociados a los objetos, reduciendo su variación a un rango especificado (normalmente entre [0, 1]), aplicando una transformación lineal. La normalización permite asignar a todos los atributos el mismo peso durante el proceso de clusterización, llevándolos a un sistema de medida adimensional.

Se recomienda aplicar la normalización de los datos antes de ejecutar el agrupamiento. No existe una regla universal para normalizar los objetos, la selección de la medida a aplicar depende del problema en particular a abordar, de las características de los atributos y de la tarea de *clustering* seleccionada. Algunas técnicas de normalización se presentan en [48 - 49].

2.2.3.1. Normalización Z-score

Se usa para llevar todos los objetos a una forma de puntuación estándar. Dado el objeto X_i con n observaciones, la estandarización *Z-score* para el atributo x_{ij} se define como la ecuación (8).

$$Z - score = Z(x_{ij}) = \frac{x_{ij} - \bar{x}_i}{\sigma_i} \quad (8)$$

Donde,

$$X_i = (x_{i1}, x_{i2}, \dots, x_{in})$$

$$\bar{x}_i = \frac{1}{n} (x_{i1} + x_{i2} + \dots + x_{in})$$

$$\sigma_i = \frac{1}{n} \left(\left| x_{i1} - \bar{x}_i \right| + \left| x_{i2} - \bar{x}_i \right| + \dots + \left| x_{in} - \bar{x}_i \right| \right)$$

Para la ecuación (8), σ_i y \bar{x}_i representan la desviación absoluta media y el valor medio del objeto X_i , respectivamente. El uso de la desviación absoluta media es más robusta frente a los *outliers* que la desviación estándar, debido a que la diferencia entre el *i-ésimo* dato y el valor medio del objeto X_i no se eleva al cuadrado. Existen otras medidas de dispersión más robustas, como la dispersión mediana, sin embargo la ventaja de usar la desviación absoluta media es que los *outliers* no se hacen demasiado pequeños, y por lo tanto, permanecen detectables [50].

2.2.3.2. Normalización *Min-Max*

Las unidades se transforman a un valor entre $[0, B]$, a partir del mínimo y máximo valor que presente el objeto X_i .

$$Z_{Min-Max}(x_{ij}) = B \times \frac{x_{ij} - x_{Min}}{x_{Max} - x_{Min}} \quad (9)$$

Donde,

$$X_i = (x_{i1}, x_{i2}, \dots, x_{in})$$

x_{Min} , es el mínimo valor del objeto X_i

x_{Max} , es el máximo valor del objeto X_i

B , Constante de variación usada para dar un rango de variación mayor a los datos normalizados

La estandarización *Min-Max* es útil cuando se trabaja con series de tiempo, por lo cual la estandarización permite controlar la variación de la magnitud entre $[0, B]$. Sin embargo, la forma de la serie de tiempo permanece invariante.

2.2.4. Criterio de agrupamiento basado en medidas de disimilaridad

Uno de los problemas más estudiados en las técnicas de agrupamiento de *clustering* ha sido la definición del criterio de agrupamiento. Es necesario aclarar que no hay un criterio de medida de similitud que pueda ser aplicado de forma universal, la selección de un criterio de agrupamiento determinado siempre se realiza de manera subjetiva [9].

La medida de similitud puede estar expresada en términos de la distancia entre objetos “disimilaridad” o en función de la “similitud” entre ellos. Elegir el criterio adecuado para evaluar la similitud entre objetos es bastante importante dentro del proceso de agrupamiento, ya que éste define el *cluster* al que se agrupa un objeto específico.

La distancia entre dos objetos X_1 y X_2 indica el nivel de disimilaridad. Entre mayor sea la distancia, mayor diferencia (disimilaridad) hay entre ellos. Con esta medida se puede evaluar el grado de semejanza entre ambos objetos, en relación a un cierto número de atributos cuantitativos.

Toda medida de distancia debe verificar al menos las siguientes propiedades:

$d(X_1, X_2) > 0$, La distancia siempre es positiva

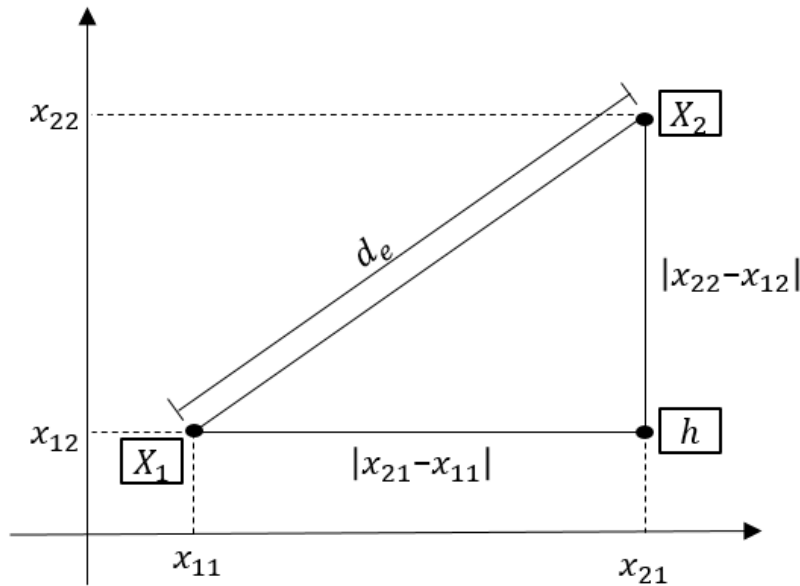
$d(X_1, X_1) = 0$, La distancia entre un objeto y el mismo siempre es cero

$d(X_1, X_2) = d(X_2, X_1)$, La distancia es una función simétrica

2.2.4.1. Distancia Euclidiana

Se deduce a partir del teorema de Pitágoras y se refiere a la distancia más corta entre dos puntos en un espacio euclídeo. Para ilustrar, a modo de ejemplo considérese el sistema de coordenadas cartesianas de espacio bidimensional de la Figura 4.

Figura 4. Deducción distancia entre dos objetos en plano cartesiano



Para el sistema de coordenadas cartesianas, la distancia Euclidiana d_e entre los puntos X_1 y X_2 es:

$$d_e(X_1, X_2) = \sqrt{|x_{21} - x_{11}|^2 + |x_{22} - x_{12}|^2} \quad (10)$$

De forma general, sean X_i y X_j dos objetos (vectores) de dimensión n , la distancia Euclidiana entre los dos objetos se calcula de acuerdo a la ecuación (11).

$$d_e(X_i, X_j) = \sqrt{|x_{j1} - x_{i1}|^2 + |x_{j2} - x_{i2}|^2 + \dots + |x_{jn} - x_{in}|^2} \quad (11)$$

Donde,

$$X_i = (x_{i1}, x_{i2}, \dots, x_{in})$$

$$X_j = (x_{j1}, x_{j2}, \dots, x_{jn})$$

2.2.4.2. Coeficiente de correlación de Pearson

Al emplearse de manera directa es considerado una medida de similaridad, ya que entre mayor sea el valor encontrado (máximo valor alcanzado es uno), mayor será la semejanza (similaridad) entre los objetos. Adicionalmente, el coeficiente de correlación de Pearson cumple con las siguientes propiedades asociadas a las medidas de similaridad:

$s(X_1, X_2) \leq 1$, La similaridad entre dos objetos diferentes nunca supera la unidad

$s(X_1, X_1) = 1$, La similaridad entre un objeto y el mismo siempre es la unidad

$s(X_1, X_2) = s(X_2, X_1)$, La similaridad es una función simétrica

Este coeficiente mide el grado de relación entre dos variables de tipo cuantitativas. Los coeficientes de correlación varían entre $[-1, 1]$, intervalo que va desde una relación negativa perfecta (-1) a una relación positiva perfecta (+1).

Sean dos vectores X y Y de dimensión n (n es una muestra que indica el número de observaciones), la correlación r producto-momento se calcula mediante la ecuación (12).

$$r = \cos(\alpha) = \frac{\sum_{i=1}^n (X_i - \bar{X}) \cdot (Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (12)$$

Donde,

$\cos(\alpha)$ Es el coeficiente de correlación muestral de Pearson

\bar{X} , \bar{Y} indican los valores medios de los vectores

Para la ecuación (12), el coeficiente de correlación r es un número absoluto adimensional y $\cos(\alpha)$ representa el ángulo entre los vectores X y Y .

Si:

$r = 1$, $\alpha = 0^\circ$, Indica que ambos vectores son colineales (alta correlación positiva).

$r = 0$, $\alpha = 90^\circ$, Indica una ortogonalidad entre los vectores (no se demuestra una correlación).

$r = -1$, $\alpha = 180^\circ$, Indica que ambos vectores son colineales de dirección opuesta (alta correlación negativa).

Con la ecuación (13) se calcula el grado de disimilaridad entre los vectores X y Y a partir del coeficiente de Pearson [50].

$$d_p(X,Y) = \frac{1-r}{2} \quad (13)$$

De la ecuación (13), si los dos vectores tienen una alta correlación positiva, el grado de disimilaridad entre ellos es cercano a cero. Por el contrario, si entre los dos vectores existe una correlación alta negativa, la disimilaridad entre ellos es cercana a 1. Para algunas aplicaciones es necesario asignar el mismo valor de disimilaridad a correlaciones altas positivas como negativas, por lo cual se puede utilizar la ecuación (14).

$$d_p(X,Y) = 1 - |r| \quad (14)$$

2.2.5. Índices de validación de *Clustering*

El agrupamiento mediante técnicas de *Clustering* es un proceso “No supervisado”, por lo cual se hace importante evaluar el resultado obtenido. Sin embargo, es difícil definir cuando un resultado es aceptable o no, ya que no tenemos para contrastar un conocimiento previo o etiqueta de la solución real del problema. Debido a esta situación, se han estudiado diferentes técnicas de validación que permiten tomar una decisión en cuanto a la calidad del resultado. Para este caso la “validación” se

relaciona con la obtención de una agrupación adecuada al problema que se quiere resolver.

Una técnica de validación interna permite elegir el mejor algoritmo de *Clustering*, así como la mejor agrupación, sin conocimiento previo de la solución y valiéndose de dos criterios que permiten evaluar la calidad de los resultados: Cohesión y Separación. Dada la naturaleza subjetiva del problema, se pueden encontrar varias agrupaciones “válidas” para un determinado número de conjuntos [51]. A manera de ejemplo, si para $k=2$ las etiquetas de agrupamiento son “Elementos Buenos” y “Elementos con fallos”, sería normal que para $k=3$ los objetos se agrupen en *clusters* más finos como “Elementos Buenos”, “Elementos con fallos Eléctricos” y “Elementos con fallos mecánicos”.

Generalmente la medida de validación se realiza con un índice que involucra en su cálculo los criterios de cohesión y separación, para determinar qué tan similares son entre si los objetos agrupados en un mismo *cluster* y que tan diferentes son respecto a los agrupados en otros *clusters*. Algunos de los índices más populares se nombran a continuación.

2.2.5.1. Índice Xie-Beni

$$S = \frac{\sum_{i=1}^k \sum_{c=1}^n (u_{ic})^2 d^2(x_i, v_c)}{n \min_{i,c} d^2(v_i, v_c)} \quad (15)$$

Donde,

$d_{\min} = \min_{i,c} d^2(v_i, v_c)$, es la mínima distancia entre Centroides del *cluster*

Para la ecuación (15), un valor pequeño de S indica que todos los *clusters* formados tienen una alta cohesión (numerador tiende a tener un valor pequeño) y están bien separados uno del otro (denominador tiende a tener un valor grande) [10].

Para el FCM con un valor de $m=2$, la ecuación (15) se generaliza mediante (16).

$$S = \frac{J(\mathbf{X}; \mathbf{U}, \mathbf{V})}{n(d_{\min})^2} \quad (16)$$

El objetivo de la ecuación (15) es minimizar el valor de S con el fin de obtener agrupamientos válidos. Para lograrlo se puede minimizar $J(\mathbf{X};\mathbf{U},\mathbf{V})$, el cual es el objetivo del FCM. Igualmente, entre mayor sea el valor de $(d_{\min})^2$ más separados estarán los *clusters* y menor será el valor de S . Es de notar que S tiende a ser decreciente para altos valores de n , lo que favorece para que el número de agrupaciones sea creciente. Para detener esto se pueden aplicar funciones de castigo *ad_hoc* estudiadas en [52].

2.2.5.2. Índice Davis-Bouldin

$$DB = \frac{1}{k} \sum_{i=1}^k \max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(v_i, v_j)} \right) \quad (17)$$

Donde,

k , es el número de *clusters*

σ_x , es la distancia media de todos los elementos del *cluster* c_x al Centroide v_x

$d(v_i, v_j)$, es la distancia entre los Centroides v_i y v_j

Para el índice, el valor $\max_{i \neq j} \left(\frac{\sigma_i + \sigma_j}{d(v_i, v_j)} \right)$ representa la peor agrupación para el clúster

i . La solución más óptima es aquella que alcanza el valor DB más bajo [53].

2.3. AGRUPACIÓN DE SERIES DE TIEMPO

Una serie de tiempo es una secuencia de valores observados y ordenados cronológicamente en una estampa de tiempo. Aunque el tiempo es una variable continua, en la práctica se toman mediciones en periodos equidistantes, por lo cual es difícil imaginar algún área científica en la que no aparezcan datos que puedan ser considerados como series temporales [50].

Las series de tiempo pueden ser de valor real o discreto, muestreadas o no uniformemente, de tipo univariantes o multivariantes y de igual o diferente longitud [54]. Se caracterizan principalmente por tener una alta dimensión y un alto nivel de ruido [55].

El estudio de una serie de tiempo se basa en descomponer las series en varias componentes [56]:

- **Tendencia:** Es la dirección hacia la que se mueve la variable en un periodo de observación. Se puede definir como el cambio general y persistente a largo plazo de la medida de la serie. La tendencia se utiliza para determinar si una serie responde en el largo plazo a una determinada ley (tendencia creciente, tendencia decreciente, tendencia estable). La tendencia se puede ajustar matemáticamente a diferentes funciones, como lineal, exponencial o parabólico. Cuando se analiza la tendencia, se prescinde de las variaciones a corto y mediano plazo. La tendencia es fundamental para el análisis de *clustering*, ya que será una medida para modelar diferentes grupos y se realizarán medidas de distancia para establecer si un elemento pertenece a un grupo a otro. Un ejemplo de esta componente sería la tendencia creciente de la demanda energética de un país, debido al auge en el sector construcción y al incremento de la explotación minera.
- **Estacionalidad:** Corresponde a fluctuaciones periódicas de la variable observada, en periodos de tiempo relativamente cortos. Se presenta debido a la influencia de fenómenos que se repiten de manera periódica. Un ejemplo se puede obtener cuando se realiza análisis del consumo de energía eléctrica de los usuarios residenciales, el cual presenta diariamente un incremento de la demanda (estacionalidad) durante las horas pico.
- **Otras fluctuaciones irregulares:** Es el residuo resultante después de extraer de las series la tendencia y las variaciones estacionales. Estos residuos pueden ser, o no, de naturaleza aleatoria. En esta componente se puede asociar el ruido producido por fenómenos de carácter ocasional aleatorio y que le imprime a las series valores atípicos. Estos tipos de fenómenos pueden ser tormentas, terremotos, huelgas, etc. Un ejemplo de estos puede ser un valor de demanda alto en una hora del día donde regularmente es bajo, debido a un eclipse solar en la zona de influencia. En algunos casos es difícil determinar la causa de la atipicidad en los datos, por lo cual es difícil de aislar en el análisis para todo el conjunto de datos.

El análisis de series temporales es fundamental en muchos campos del conocimiento y han ganado atención en temas de investigación sobre agrupamiento, búsqueda de similaridad, clasificación, predicción, entre otras [56]. Se puede pensar que con una comparación secuencial de dos series temporales bastaría para determinar si son similares o no. Sin embargo, no es suficiente con realizar un rastreo de dos secuencias temporales para extraer conocimiento de una base de datos de altas dimensiones.

El agrupamiento de series de tiempo ha sido objeto de múltiples investigaciones desde diferentes campos, entre los que se destacan investigaciones financieras, la medicina, la geología, la meteorología, la bioquímica, la ecología, la genética, entre otros [57]. Fruto de estas investigaciones se han propuesto desde la minería de datos, diferentes algoritmos y metodologías para su agrupamiento [58].

Muchos de los algoritmos de agrupación de series de tiempo tienden a convertir los datos de series temporales en una forma de datos estáticos para que los algoritmos de agrupación tradicionales puedan ser usados directamente. Otros, adecuan los métodos tradicionales para que se utilicen directamente con las series de tiempo [54]. La estructura de las series de tiempo hace que muchos métodos de agrupación tradicionales no se puedan aplicar de manera directa.

El agrupamiento de series de tiempo se puede abordar de dos maneras [58 - 59]:

- Basado en características o basado en el modelo (*feature-based* o *model-based*). Es decir, resumiendo o transformando previamente los datos iniciales por medio de extracción de características o modelos paramétricos. Por ejemplo, Regresión dinámica, ARIMA, redes neuronales. De esta forma se transforman los datos a un espacio de menor dimensión, en el cual las técnicas de *Clustering* trabajan directamente. Los problemas que puedan presentar se asocian con una pérdida de información de la serie por la representación de la misma por medio de características o modelos.
- Basado en datos “crudos” (*raw-data-based*). Las técnicas de agrupamiento se aplican directamente sobre los vectores de series de tiempo sin realizar ninguna transformación espacial previa. Al trabajar directamente sobre la serie previenen la pérdida de información y es flexible para tratar longitudes de datos de series variables. La dificultad de estos algoritmos es que trabajan con espacios de alta dimensión, lo cual puede sacrificar la calidad de la solución; adicional, tienen alta exposición y sensibilidad al ruido en los datos de entrada. Varios ejemplos se pueden encontrar en [57].

Agrupar series de tiempo con extensos puntos de muestra (largas series) hace que el proceso de *Clustering* sea lento, debido a la alta dimensionalidad del problema, lo que causa sesgos en el proceso de agrupación, ya que las diferencias entre objetos “similares” y la diferencia entre objetos “diferentes” puede llegar a ser confusa, generando dificultades en la interpretación de los grupos formados [60].

El propósito de agrupamiento de series de tiempo es resaltar la estructura inherente en el conjunto de datos, agrupándolos en grupos homogéneos, de forma que la similaridad entre series en un mismo *cluster* sea máxima. El resultado obtenido de este proceso puede revelar comportamientos o tendencias comunes que ayudan a un mejor entendimiento de los datos analizados [58].

2.4. ALGORITMOS GENÉTICOS

La evolución hace referencia a cualquier proceso de cambio en el tiempo. Los algoritmos genéticos (AG) son un tipo de algoritmos evolutivos, usados para resolver problemas de búsqueda y optimización. Se sustentan en la teoría de la evolución de Charles Darwin, donde los individuos más adaptados al medio tienen

mayor probabilidad de sobrevivir, reproducirse y conservar sus mejores características por varias generaciones. Este proceso evolutivo no debe ser del todo determinístico y en cierto grado posee una naturaleza estocástica, dadas las eventualidades de la naturaleza que puedan causar la desaparición de individuos bien dotados, ya sea por competencia entre sí o por la aparición de depredadores imprevistos [61].

Los AG son una buena opción para encontrar soluciones de calidad, cuando se presenta alguna de las siguientes situaciones [62]:

- No hay un algoritmo exacto para resolver el problema.
- El problema es flexible, es decir, acepta soluciones de buena calidad sin importar que sea la óptima.
- Los datos usados para la solución son imprecisos o probabilísticos.
- Limitaciones por tiempo, espacio y/o memoria.
- Se requiere encontrar una solución rápida y usarla como punto de partida para un método exacto.
- Se desea explorar de manera amplia y eficiente el espacio de soluciones.

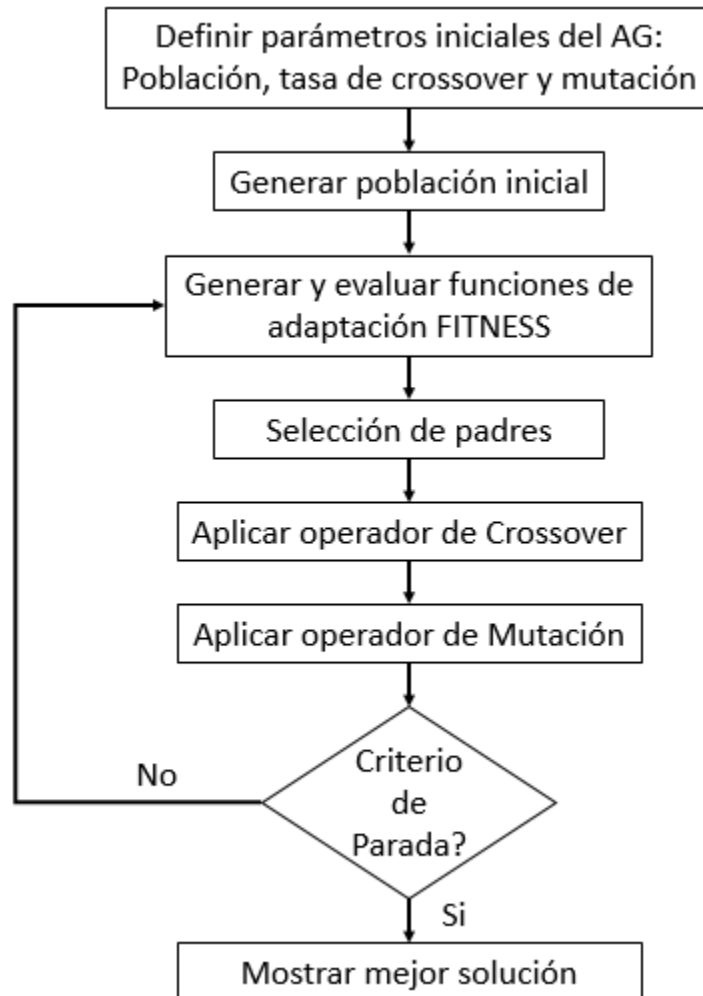
Los AG nacen de la mano de John Holland en los años 70 [14], sin embargo hasta los 90 se hicieron populares para la solución de diversos problemas de la vida real [63]. Se presentan de manera iterativa como una secuencia de pasos inspirados en la teoría de la evolución, en la cual interactúa una población de individuos de igual especie y de operaciones genéticas entre sus cromosomas, de tal manera que los más adaptados al medio tengan mayor probabilidad de sobrevivir al proceso natural de selección, consiguiendo que sus genes sean conservados por generaciones.

Con el fin de hacer evolucionar una población inicial y obtener en cada generación una solución de mejor calidad, los AG en su forma canónica se compone de las siguientes etapas:

- Selección.
- Recombinación o cruzamiento (*crossover*).
- Mutación.

La Figura 5 muestra el diagrama de flujo para un algoritmo genético básico [62].

Figura 5. Esquema algoritmo genético básico



Uno de los mayores desafíos a los que se enfrentan los AG es la codificación del problema, puesto que la evolución de las soluciones parcialmente encontradas hacia las soluciones óptimas del problema, depende en buena medida de una adecuada representación de las mismas. Esto se debe a que los operadores genéticos son modificados de acuerdo al sistema utilizado para la codificación. La codificación binaria es la más utilizada, sin embargo, una práctica sana es determinar la naturaleza de las variables del problema y emplear por comodidad una codificación de igual naturaleza [62].

Cada iteración del algoritmo de la Figura 5 se conoce como un ciclo generacional, y cada solución alcanzada en la nueva generación se conoce como cromosoma. Los cromosomas evolucionan en cada generación a través de los operadores genéticos *Crossover* y *mutación*, y se evalúa su aptitud por medio de una función

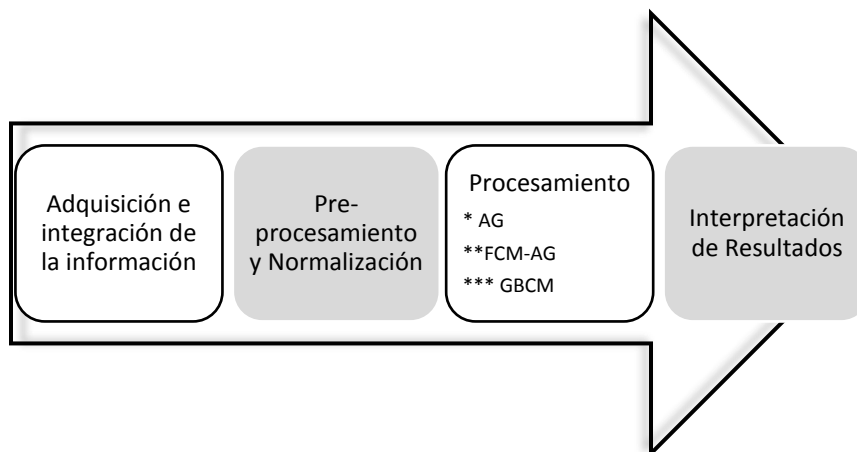
fitness. En cada generación se aplica repetida y sucesivamente los operadores genéticos hasta alcanzar un criterio de parada que generalmente se puede determinar por un número de ciclos generacionales dado, por alcanzar una solución con una calidad mínima especificada, por pérdida de diversidad de los cromosomas entre cada generación o por una combinación de las anteriores.

No hay certeza para afirmar que la solución alcanzada mediante el AG sea la solución global del problema, incluso no se puede afirmar que la solución alcanzada sea de buena calidad, esto depende de la codificación adecuada del problema y del correcto dimensionamiento de la población inicial. La población inicial debe garantizar un adecuado nivel de diversidad entre los individuos tal que el algoritmo cada vez explore áreas diferentes del espacio de búsqueda. El tamaño de la población inicial determinará la cantidad de soluciones a manejar en todo el proceso, lo que es un factor importante para la convergencia.

3. METODOLOGÍA PROPUESTA

La Figura 6 presenta el esquema metodológico general de la metodología aplicada, la cual se desarrolla en cuatro etapas principales.

Figura 6. Esquema metodológico general



Las dos primeras etapas de la Figura 6 (Adquisición e integración de la información y Pre-procesamiento y Normalización) son etapas comunes a cualquier proyecto de minería de datos. La etapa de procesamiento corresponde al núcleo principal del trabajo de grado, en la cual se aplican tres metodologías para la agrupación de series de tiempo correspondientes a balances de energía mensuales en transformadores de distribución (pérdidas de energía mensuales por transformador). Los resultados son interpretados en la etapa final. Las tres metodologías aplicadas durante la etapa de procesamiento son:

1. Algoritmo Genético (AG)
2. Metodología híbrida *Fuzzy C-Means* – Algoritmo Genético (FCM-AG)
3. Metodología jerárquica divisiva *Bisecting Fuzzy C-Means* (GBCM)

3.1. ADQUISICIÓN E INTEGRACIÓN DE LA INFORMACIÓN

La información utilizada para el desarrollo del presente trabajo se obtiene a partir de consultas a una base de datos del sistema comercial de una empresa de distribución de energía en Colombia. La información se extrae para un periodo histórico de $n = 34$ meses y corresponde a:

1. Suma de consumo de energía facturada a usuarios, desagregada por transformadores de distribución.
2. Consumo de AP y otras cargas aforadas en cada uno de los transformadores.
3. Consumo de energía registrado en macromedidores.

Con el fin de obtener las pérdidas totales por transformador, se realizan balances de energía a cada uno de ellos aplicando la ecuación (2) durante un periodo histórico de 34 meses. Los balances de energía se organizan en forma de vector en una matriz de dimensiones $N \times n$, la cual tiene la siguiente estructura:

$$\begin{bmatrix} \text{Trafo1} \\ \text{Trafo2} \\ \vdots \\ \text{TrafoN} \end{bmatrix} = \begin{bmatrix} \text{Perd_mes1_Trafo1} & \text{Perd_mes2_Trafo1} & \dots & \text{Perd_mesn_Trafo1} \\ \text{Perd_mes1_Trafo2} & \text{Perd_mes2_Trafo2} & \dots & \text{Perd_mesn_Trafo2} \\ \vdots & \vdots & \ddots & \vdots \\ \text{Perd_mes1_TrafoN} & \text{Perd_mes2_TrafoN} & \dots & \text{Perd_mesn_TrafoN} \end{bmatrix}$$

Donde,

n : Número de meses de observación o atributos por transformador. 34 meses.

N : Número de transformadores u objetos. 7616 transformadores.

La matriz se almacena en un archivo tipo .x/sx con el fin de importarla posteriormente a Matlab®.

La caracterización de los balances de energía almacenados se muestra en la Tabla 1.

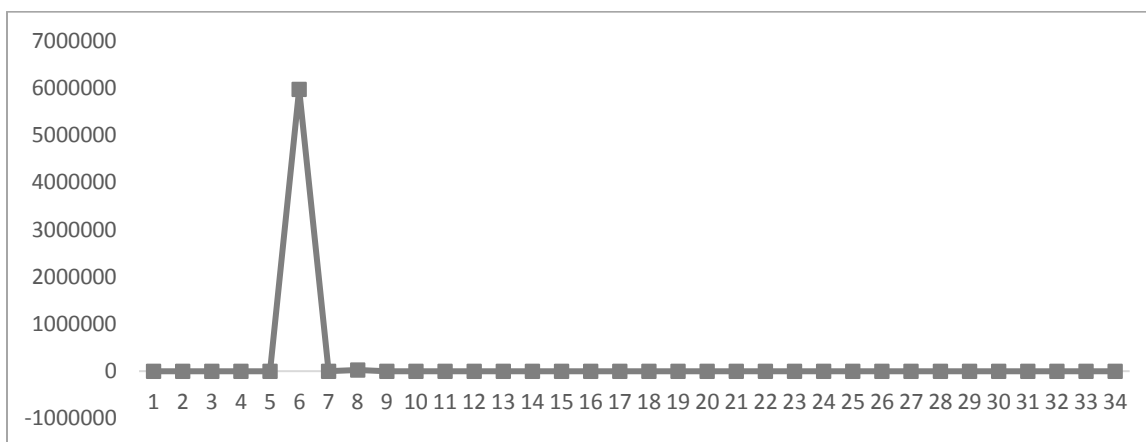
Tabla 1. Caracterización balances iniciales

Periodo de análisis (meses)	34
Número de objetos	7616
Total de atributos	258.944
Valor medio de los datos [kWh/mes]	146
Valor mínimo de los datos [kWh/mes]	- 108.937*
Valor máximo de los datos [kWh/mes]	5.974.465**
Desviación Estándar	11.857

Es de notar en la Tabla 1 que existen valores que por su magnitud se pueden identificar a simple vista como datos atípicos. Es el caso del valor mínimo (-108.937* kWh/mes) y el valor máximo (5.974.465** kWh/mes), los cuales se encuentran muy desviados respecto al valor promedio (146 kWh/mes). Estos valores atípicos se pueden presentar por diferentes factores externos, entre los cuales se cuentan: errores humanos asociados a la toma de lectura del macromedidor o usuarios, por daños en los medidores y/o por sucesos estocásticos en el proceso de facturación, como por ejemplo el cobro en un mes puntual de energía adicional a un usuario por concepto de una recuperación de energía.

Para ilustrar mejor la ocurrencia de valores atípicos se detalla el caso del valor máximo. Revisando el historial de facturación del transformador se identifica que sobre el macromedidor se presenta un daño en el fichero que da la lectura, debido a esto en el mes 6 la lectura es inferior a la tomada en el mes 5. Esta situación no es normal, ya que la lectura en un medidor para un mes posterior debe ser superior. Durante el proceso de facturación se interpreta como si el medidor hubiese completado una vuelta completa del fichero y volviera a iniciar su contador, por tanto la energía facturada en el macromedidor para el mes 6 es muy superior al promedio de facturación mensual. Por lo anterior, el resultado del balance de energía para el mes 6 da como resultado unas pérdidas desproporcionadas y muy alejadas del valor medio. En la Figura 7 se ilustra el comportamiento histórico de las pérdidas para el transformador en asunto.

Figura 7. Valor atípico de pérdidas



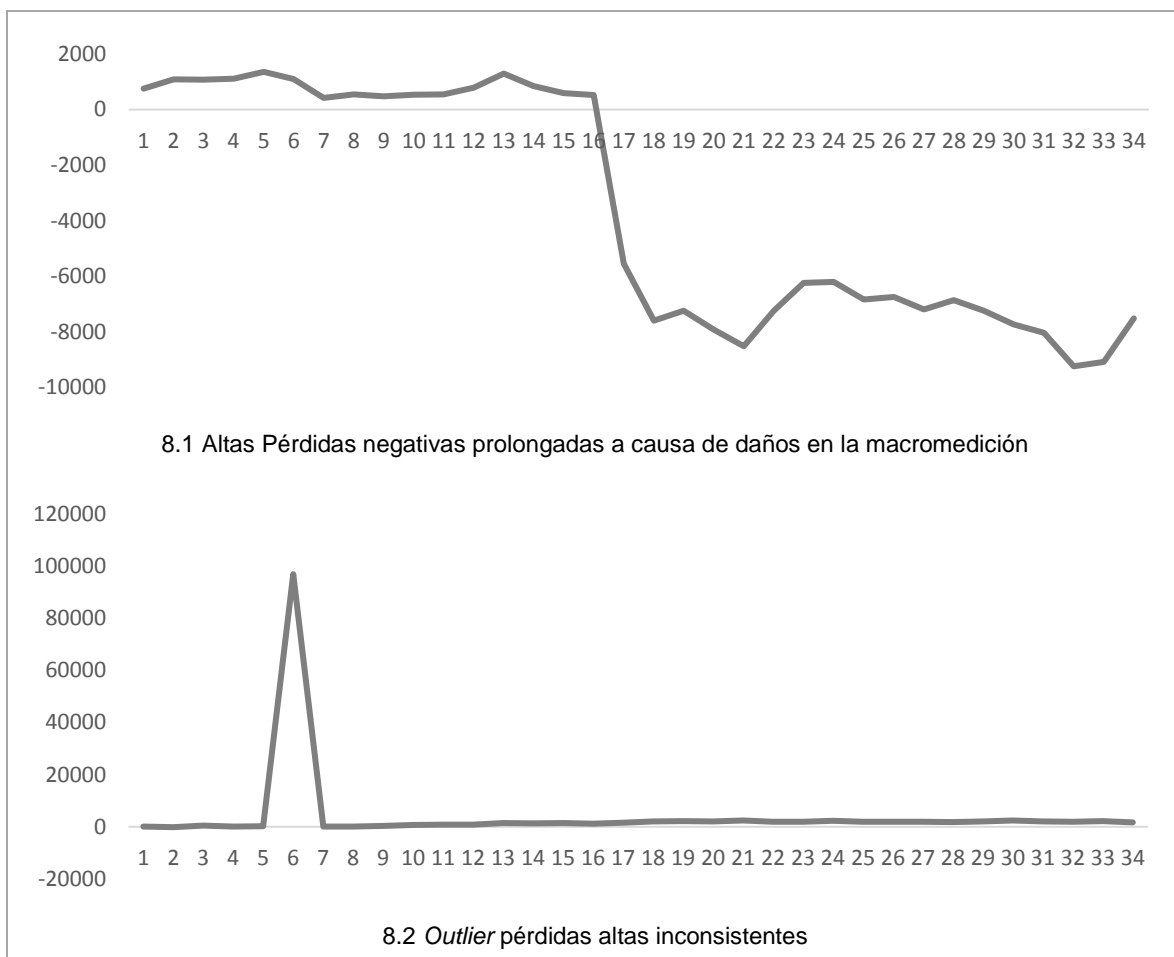
De la Figura 7 se observa que las pérdidas en el transformador son constantes y de baja magnitud. A simple vista se identifica en el mes 6 un dato atípico muy alejado de la media mensual. Este comportamiento muestra el valor agregado que a través de la crítica post-facturación de las empresas se le puede dar al proceso de limpieza de la bases de datos, lo cual será muy útil al momento de aplicar estrategias de minería de datos para encontrar tendencias.

Debido a los casos atípicos presentados en la base almacenada, se requiere de un proceso previo de pre-procesamiento de los datos, en el cual se identifiquen y se corrijan los casos particulares identificados.

3.2. PRE-PROCESAMIENTO Y NORMALIZACIÓN DE LOS DATOS

El pre-procesamiento de los datos se realiza con el fin de identificar datos ausentes, erróneos, atípicos, que a criterio experto son redundantes o no significativos para el análisis y pueden inferir errores en la metodología de minería de datos a aplicar. Existen dos opciones a seguir en pre-procesamiento cuando se identifican estos datos: la primera opción es remover los objetos para realizar un análisis de manera independiente. En este caso el espacio de características se comprime de R^p a $R^{p'}$, donde $p' < p$. La segunda es corregir el valor fuente del error. La elección de una de las dos alternativas se toma a partir del conocimiento previo del analista sobre el problema y los datos.

Figura 8. Casos típicos filtrados de acuerdo a un análisis de pre-procesamiento



De acuerdo a la Figura 8, los objetos clasificados en los casos tipo 8.1 se removieron de la base de datos inicial debido al error que pueden generar al método de agrupamiento o porque de acuerdo a la experiencia del analista no son significativos para identificar pérdidas no técnicas en los transformadores. Para los casos tipo 8.2 se corrigieron los datos atípicos de acuerdo a índices estadísticos (mínimo, máximo, media, desviación estándar).

Revisando más detalladamente la base de datos, se encuentra que los casos tipo 8.1 con pérdidas negativas inferiores a -500 kWh/mes se presentan por daños en el macromedidor, lo que valida la hipótesis de que no aportan en la identificación de pérdidas. Para el analista de la empresa de distribución estos casos se denominan balances de energía “no confiables” y merecen un tipo de análisis diferente, por lo cual el direccionamiento de los transformadores en campo se realiza como una reparación del macromedidor.

Los casos tipo 8.2 con pérdidas superiores a 20.000 kWh/mes se presentan principalmente por errores asociados al proceso de facturación del macromedidor. Estos valores atípicos se corrigen a partir del comportamiento histórico de la serie analizada.

Una vez se ejecuta el pre-procesamiento en la base de datos inicial, los datos a analizar tienen las siguientes características:

Tabla 2. Caracterización datos iniciales después de etapa de pre-procesamiento

Periodo de análisis (meses)	34
Número de objetos	5125
Total de atributos	174.250
Valor medio de los datos	195
Valor mínimo de los datos	-500
Valor máximo de los datos	19.442
Desviación Estándar	523

Ya con los datos pre-procesados se pasa a la etapa de normalización. A cada serie de tiempo o vector se le aplica la normalización $Z_{Min-Max}$ revisada en la sección 2.2.3.2, asumiendo un valor de $\beta = 10$. Con la normalización elegida se garantiza un rango de variación de la magnitud de las series entre [0, 10] sin afectar la forma de las mismas.

3.3. PROCESAMIENTO DE LOS DATOS

Esta etapa de procesamiento se enmarca en el área de minería de datos, principalmente en la aplicación de técnicas para el agrupamiento de series de tiempo. Se aplican tres metodologías, la primera es un algoritmo genético puro actuando como un agrupador de datos. La segunda corresponde a un algoritmo híbrido entre los métodos *Fuzzy C-Means* y algoritmo genético (FCM-AG). La tercera es a una adaptación del *Bisecting k-Means*, nombrada en este trabajo como Algoritmo genético divisivo *Bisecting C-Means* (GBCM).

Los objetos a agrupar son los balances de pérdidas históricos de los transformadores organizados en forma de serie de tiempo, como se ilustra en la Figura 1. Los resultados obtenidos son transformadores agrupados en *clusters*, de los cuales se pueden determinar patrones de comportamiento de sus balances en el tiempo, permitiendo inferir “comportamientos anómalos” que suponen pérdidas no técnicas de energía.

Los métodos de análisis aquí propuestos se basan en datos crudos (*raw-data-based*), ya que se aplican directamente sobre las series sin realizar extracción de características previas. Trabajar con datos crudos permite formular hipótesis directas sobre los resultados obtenidos sin necesidad de devolverse a los datos iniciales, adicional que minimiza el riesgo de pérdida de información asociada a la representación de los datos por medio de características o modelos.

A partir de los resultados obtenidos se identifican transformadores en los cuales se infiere un comportamiento “sospechoso”, los cuales fueron direccionados para inspección en campo.

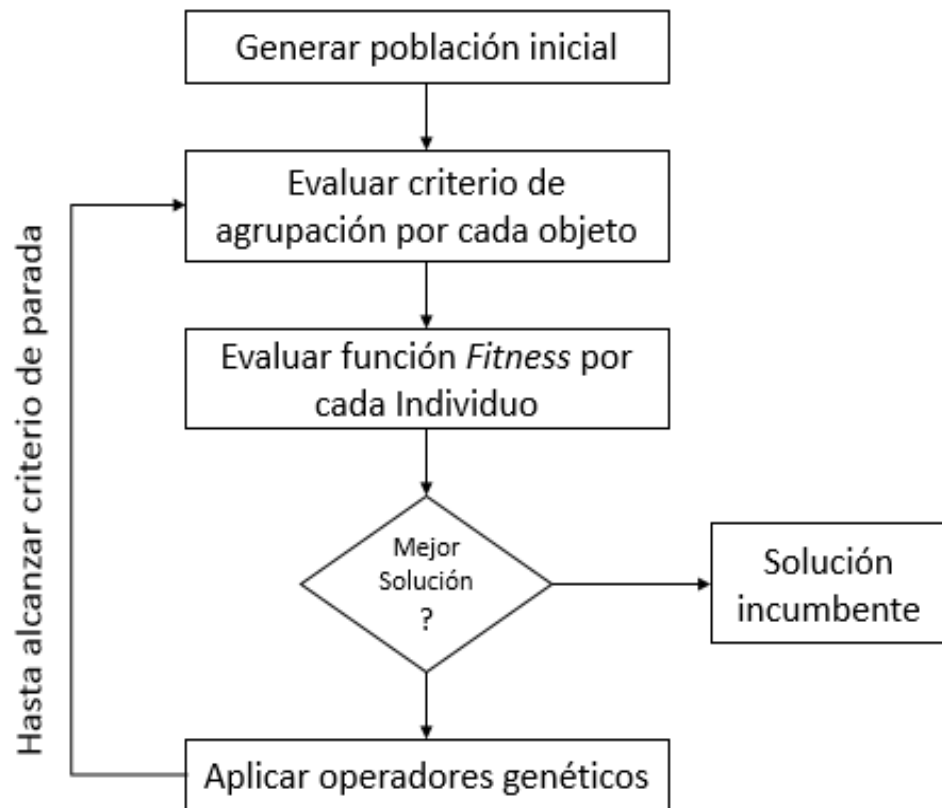
3.3.1. Agrupación mediante Algoritmo Genético (AG)

El objetivo del AG como método de agrupación es encontrar un número de *clusters* k , en el cual se pueden agrupar los objetos a analizar, garantizando que los miembros de un mismo *cluster* sean semejantes entre sí, pero diferentes a otros clasificados en diferentes *clusters*.

El diagrama de la Figura 9 muestra el esquema general de aplicación del AG.

Lo primero a realizar antes de iniciar el algoritmo genético es establecer el tipo de codificación en el cual se representarán las soluciones del problema. Para este caso, debido a la naturaleza de las variables del problema, se decide una codificación real.

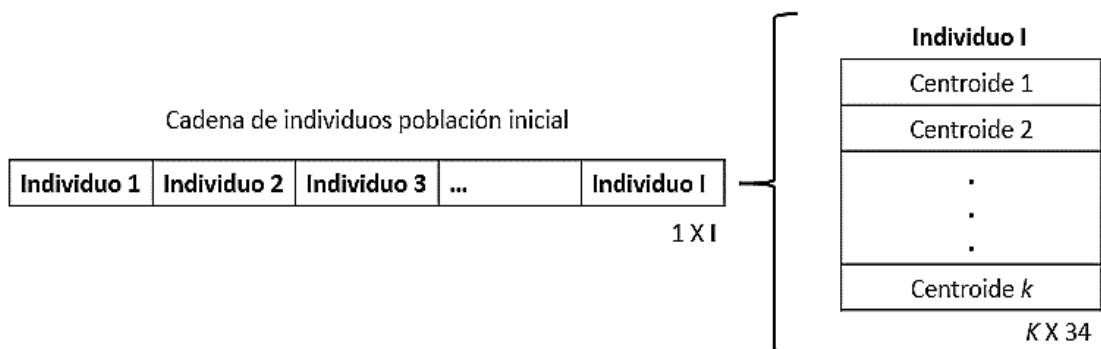
Figura 9. Diagrama esquemático del AG



3.3.1.1. Generación población inicial

La codificación de la población inicial se realiza de acuerdo como se muestra en la Figura 10.

Figura 10. Población inicial del AG

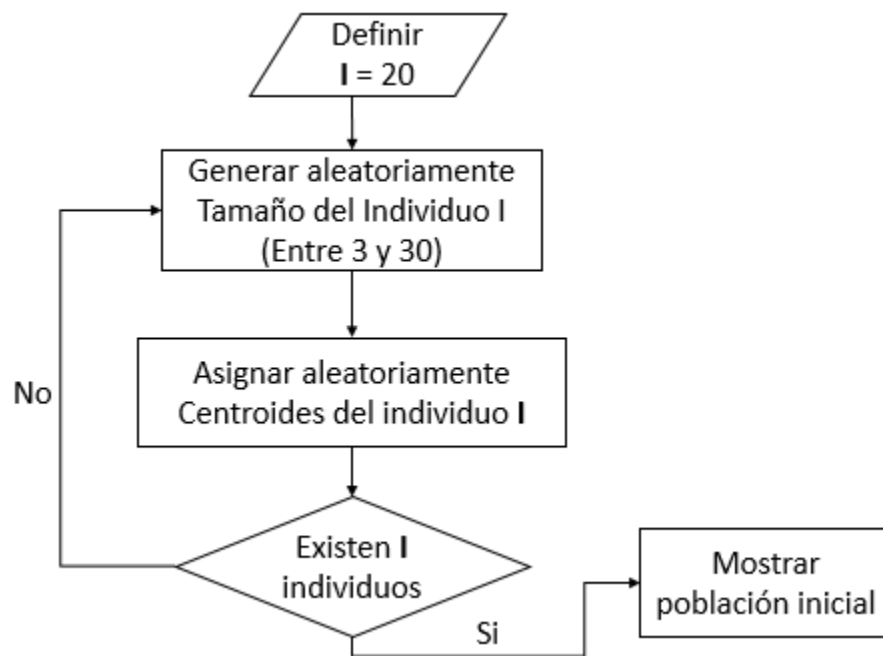


La población inicial se representa como una cadena de I individuos candidatos a ser la solución del problema¹⁷. Para conformar los I individuos, se seleccionan de forma aleatoria k series de la base de datos inicial¹⁸. Cada serie seleccionada representa un centroide y el número total de series que componen a un individuo representa el número total de *clusters* k en el que se agrupan los datos.

La longitud de cada individuo I (número de centroides por individuo) se determina de manera aleatoria, con un rango de variación $3 \leq k \leq 30$. La longitud mínima se fija en 3 para evitar que los datos se agrupen en pocos *clusters*, formando centroides muy distante de las series que lo conforman (*black holes*). La longitud máxima se fija en 30 para garantizar al AG un espectro amplio de exploración sobre el espacio de búsqueda, sin embargo no se toma un valor mayor para evitar que se formen muchos *clusters* con pocas series (*dust clouds*), castigando la generalidad en la agrupación [64 - 65].

La Figura 11 muestra el esquema de generación de la población inicial para el AG.

Figura 11. Esquema de asignación de población inicial



El proceso de la Figura 11 se repite de manera iterativa hasta formar el total de individuos I .

¹⁷ Cadena de cromosomas tamaño I

¹⁸ k Genes por cromosoma

El tamaño de la población inicial se fija en 20, valor establecido a partir de una prueba de parametrización previa del AG en la que se evalúa la calidad de la solución encontrada y el tiempo computacional empleado al variar el número de individuos I . La población inicial debe ser diversa, permitiendo una exploración amplia del espacio de búsqueda sin caer en una convergencia prematura del algoritmo.

Como restricciones en la generación de la población inicial se encuentran:

1. Los individuos candidatos a la solución varían el número de centroides en un rango de 3 a 30 centroides.
2. Para el individuo I no se debe repetir un mismo Centroide como candidato de la solución. En caso de presentarse sería una solución no factible que retrasa la convergencia del método.

3.3.1.2. Criterio de agrupación

El criterio de agrupación permite definir para cada serie u objeto de análisis el *cluster* en el cual se agrupa. Para cada uno de los I individuos seleccionados en el paso anterior se propone una agrupación de las series.

El criterio de agrupación se define en función de la “disimilaridad” de las serie a agrupar, con respecto a los centroides candidatos de solución. Las series se agrupan en el *cluster* donde encuentre mayor semejanza con el centroide. Como medida de “disimilaridad” se utilizó la distancia euclidiana de la ecuación (11).

3.3.1.3. Evaluación función *Fitness*

La función *fitness* simula la adaptación que tienen los individuos al problema y la probabilidad de que sean seleccionados posteriormente para reproducirse. Para simular este proceso en el AG se adapta una función que evalúa la calidad de la solución propuesta por cada individuo.

La función *fitness* utilizada en este proyecto corresponde al índice Davies-Bouldin (DB) presentada en la ecuación (17). Como caso particular, para medir la distancia entre centroides se utilizó el coeficiente de correlación de Pearson de la ecuación (13). De esta manera, la calidad del agrupamiento se evalúa en función de la tendencia que tienen las series clasificadas en un mismo *cluster*.

Como solución incumbente¹⁹ se selecciona el individuo I que alcanza el menor valor de la función *fitness*. En cada iteración del algoritmo se verifica la calidad de la

¹⁹ Mejor solución parcial encontrada en el desarrollo del algoritmo

agrupación de cada individuo, si se observa en alguno una calidad mejor a la de otro ya seleccionado previamente, este entrará como nuevo incumbente, desplazando el anterior. Cuando el algoritmo alcance el criterio de parada definido la última solución incumbente obtenida pasará a ser la solución final del problema.

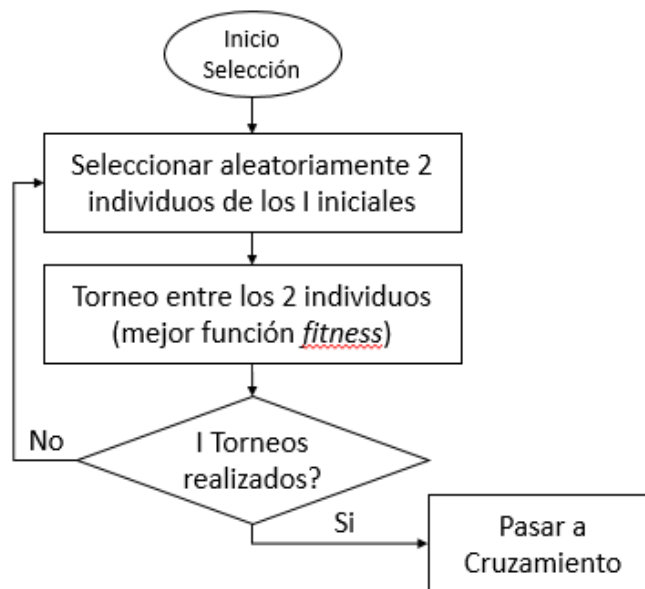
3.3.1.4. Operadores genéticos

Los operadores genéticos permiten generar diversidad genética en la población, con el fin de que a través del proceso iterativo se explore de manera amplia el espacio de soluciones, hasta alcanzar la solución más adaptada al problema.

Los operadores genéticos son: Selección, recombinación y mutación. Estos operadores se deben ejecutar en el orden mencionado con el fin de generar la población de la nueva generación.

Selección de padres: En este trabajo específico el proceso de selección se realiza mediante torneos. Se realizan I torneos, de los cuales saldrán igual número de ganadores (Padres). A cada torneo ingresa dos individuos seleccionados aleatoriamente y el ganador será aquel con mejor función *fitness*. La Figura 12 ilustra el proceso de Selección.

Figura 12. Esquema selección por torneo



El esquema de selección por torneos es utilizado debido al buen desempeño que presenta en otras investigaciones [66]. Adicionalmente, permite mejorar el rendimiento computacional del algoritmo.

Recombinación: Una vez finaliza el proceso de selección, se procede a la recombinación o cruce. Una recombinación consiste en escoger dos de las configuraciones con derecho a generar descendientes (padres) y proceder a cruzar esas configuraciones para producir dos nuevas candidatas (descendientes). Las parejas de padres se conforman aleatoriamente, evitando que una pareja de padres a cruzar se encuentre conformada por un mismo individuo.

Cada pareja seleccionada se recombina, generando igual número de descendientes, los cuales pasarán a ser parte de la nueva generación. El cromosoma del primer descendiente tiene un segmento del padre 1 y otro segmento del padre 2, el segundo descendiente corresponde al complementario genético.

El proceso de recombinación elegido es de tipo *Multi point Crossover* [61], donde los puntos de cruzamiento se obtiene generando dos número aleatorio entre 1 y la longitud total del cromosoma (para este caso particular $n=34$).

Al proceso de recombinación se asigna un operador conocido como tasa de *crossover* (ρ_c), el cual contribuye a la decisión si la pareja de padres seleccionada se cruza o no. Si la tasa de *crossover* se sobrepasa dado un evento aleatorio, el cruce genético de los individuos se lleva a cabo. Como tasa de *crossover* se selecciona un valor elevado, generalmente entre 0.7 y 1, para que las soluciones se recombinen continuamente generando mayor diversidad en las poblaciones y dando lugar a una mejor convergencia [62]. Para este trabajo se selecciona una tasa de *crossover* del 80%.

Mutación: El objetivo de la mutación es mantener diversidad genética entre cada generación, permitiendo que el algoritmo escape de óptimos locales. Se utiliza para desplazar las zonas de búsqueda de soluciones hacia zonas del espacio que no pueden ser alcanzadas por otros operadores.

La mutación consiste en alterar el valor de una o varias posiciones de la cadena cromosómica de un individuo. La mutación debe guardar un equilibrio determinístico-probabilístico.

Existen diversas maneras de mutación, en este trabajo se implementa mutación por población, ya que se facilita la aplicación dadas las características de codificación de la población inicial, adicional que permite mejorar el tiempo de cómputo [62].

Para la mutación implementada, la población inicial se toma como un cromosoma de longitud $Long = I \times k$, y el número de mutaciones a realizar se obtienen del producto de $\rho_m \times Long$, donde ρ_m es la tasa de mutación. Para el algoritmo implementado se asume una tasa de mutación $\rho_m = 10\%$ con el fin de generar diversidad genética y evitar que el proceso evolutivo se ralentice²⁰. La posición a mutar (individuo) se elige a partir del lanzamiento de un número aleatorio entre 1 y

²⁰ Hacer lento el proceso

I, el individuo seleccionado tiene cuatro posibilidades de mutación, dos de tipo determinístico y dos de tipo probabilístico. La elección de cada tipo de mutación depende de un nuevo evento aleatorio y cada una tiene igual posibilidad de ser seleccionada. Las opciones son las siguientes:

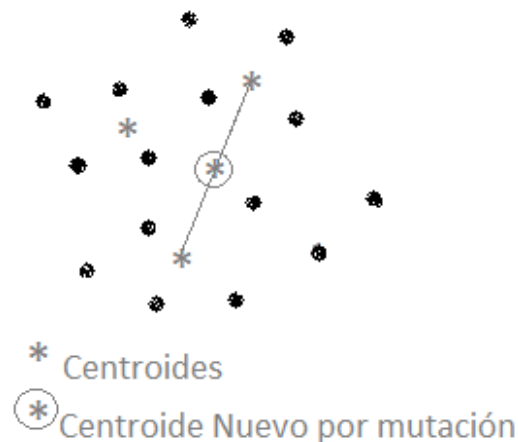
- Aumentar de manera aleatoria en uno los centroides candidatos a solución. Se aumenta en uno el número de centroides, el nuevo centroide es seleccionado aleatoriamente de la base general de series.

Figura 13. Opción uno de mutación



- Aumentar en uno los centroides candidatos a solución, a partir de los centroides más alejados. Se seleccionan los dos Centroides más alejados y se aumenta en uno el total de centroides del individuo a partir del promedio de los dos seleccionados.

Figura 14. Opción 2 de mutación



- Eliminar un Centroide aleatoriamente. De los centroides del individuo seleccionado para mutación, se elige aleatoriamente uno para ser eliminado de los candidatos a solución

Figura 15. Opción 3 de mutación



- Eliminar un Centroide, a partir de los centroides más cercanos. Se buscan los dos centroides más cercanos y elimino uno de ellos de manera aleatoria.

Figura 16. Opción 4 de mutación



3.3.1.5. Criterio de parada

El proceso de selección, cruzamiento y mutación se realiza en cada ciclo generacional conservando el orden descrito. Este proceso es iterativo, la nueva generación de individuos es sometida al proceso evolutivo, generando nuevos

individuos o candidatos de solución del problema. En cada generación se propaga buenas características genéticas hasta alcanzar un criterio de parada definido.

$$Total_Iteraciones \geq 100 \quad (18)$$

Dos de los procesos que más demandan esfuerzo computacional en el problema de agrupación mediante AG es la evaluación de la función *fitness* y evaluar el criterio de agrupación por cada objeto a agrupar. Teniendo en cuenta que en cada iteración cambia el individuo candidato de solución y que deben repetirse los dos pasos mencionados, se establece un criterio de parada basado en el total de iteraciones del algoritmo.

3.3.2. Agrupación mediante metodología híbrida FCM-AG

En la metodología híbrida el agrupamiento es visto como un problema de optimización, donde se buscan los centroides y el número de *clusters* que optimizan unas medidas de calidad pre-definidas (distancias *inter-cluster*²¹ e *intra-cluster*²²).

El FCM-AG sigue el mismo diagrama esquemático del agrupador AG, la única diferencia se encuentra en la evaluación del criterio de agrupación por cada objeto. Para este caso en vez de aplicar una función de distancia como único criterio de agrupación, se implementa el proceso de *clustering* iterativo *Fuzzy C-Means* (FCM).

En el híbrido propuesto, el AG se encarga de explorar el espacio de búsqueda global y proponer condiciones iniciales de centroides y *clusters* candidatos de solución. El FCM realiza una exploración local, a partir de las condiciones iniciales dadas por el AG, y propone una solución de agrupación por cada condición inicial. La mejor solución encontrada se guarda como solución incumbente, hasta cumplir un criterio de parada y convertirse en la solución final del problema. En la Figura 17 se muestra el diagrama esquemático del FCM-AG.

Con el diagrama esquemático de la Figura 17 se aprovecha bajo un mismo algoritmo la capacidad de búsqueda global del AG, con la explotación del espacio local del FCM, alcanzando una agrupación natural de los objetos sin requerir una etiqueta previa para ello.

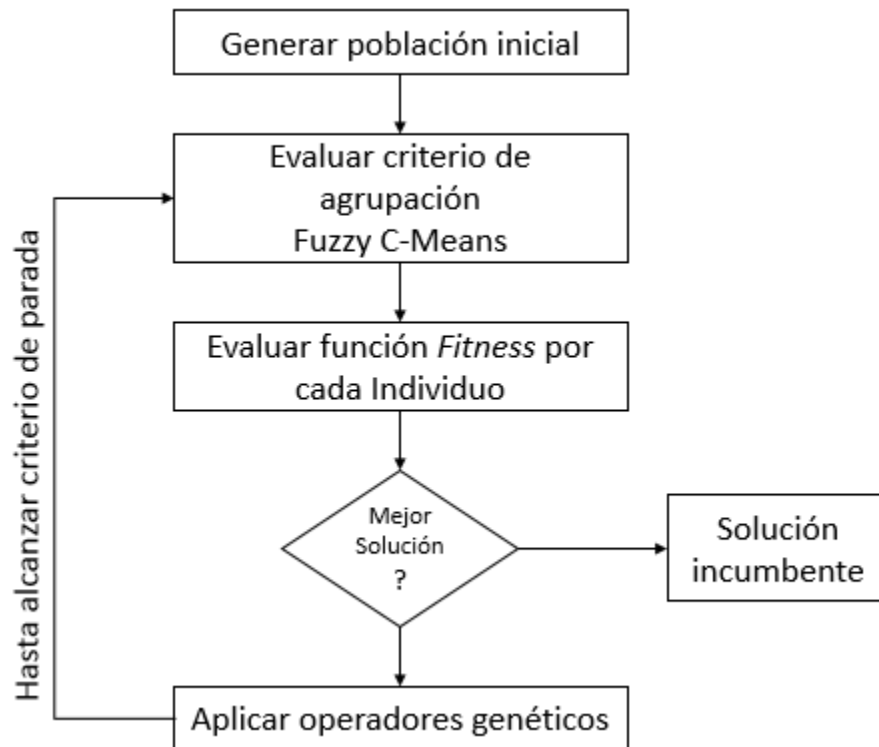
Algo importante a resaltar en esta metodología es que la partición de los datos no es dura (*Hard*), sino difusa, por lo cual para cada serie agrupada es posible conocer el porcentaje de pertenencia de esta a cada uno de los *clusters* generados.

²¹ Distancias entre los centroides de los diferentes *clusters*

²² Distancia entre cada uno de los objetos y el centroide pertenecientes a un mismo *cluster*

La parametrización del AG para la metodología híbrida es la misma que se utiliza en el AG como clasificador puro. No se requiere parametrizar nuevamente el AG.

Figura 17. Diagrama esquemático del FCM-AG



Como parametrización adicional, para el FCM se establece un valor de m igual a 2, dado el trabajo previo y las recomendaciones realizadas en [39]. El criterio de parada para el FCM se establece en función de la variación en cada iteración i de la matriz de pertenencia difusa $|\mathbf{U}^i| - |\mathbf{U}^{i-1}| < 1 \times 10^{-5}$.

Ya que la evaluación del criterio de agrupación no se realiza de manera directa con una función de distancia como en el AG, sino mediante un método iterativo como lo es el FCM, es posible que se vea afectado el tiempo de cómputo de la metodología híbrida con respecto a la agrupación mediante el AG puro.

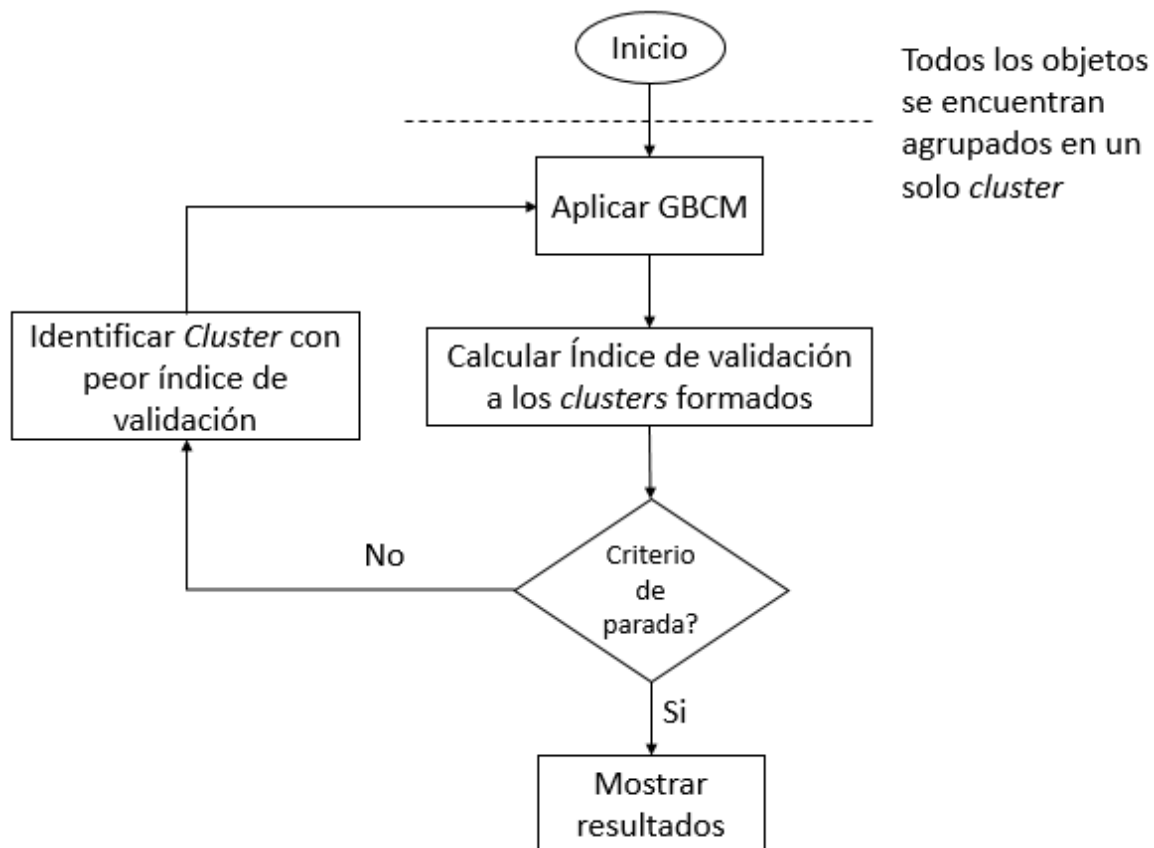
3.3.3. Método genético divisivo *Bisecting Fuzzy C-Means* (GBCM)

El GBCM es una adaptación del método *Bisecting k-Means* (B-KMA) propuesto por Steinbach et. Al [22], con las siguientes diferencias:

- El algoritmo de agrupación usado en cada etapa jerárquica es el *Fuzzy C-Means*. Este enfoque permite definir de manera difusa los centroides representativos de cada grupo (adaptabilidad) y tiene en cuenta para la agrupación los detalles dinámicos que caracterizan el patrón de las series de tiempo (sensibilidad). Así, una misma serie puede agruparse en diferentes *clusters* a partir del grado de pertenencia que la caracteriza [60].
- En cada etapa jerárquica, el número de *clusters* formados no son consecuencia de fijar un valor de $k=2$, como se hace en el caso del B-KMA. Para el GBCM el número de *clusters* obtenido en cada etapa jerárquica se obtiene de manera libre aplicando el FCM-AG. De esta manera se garantiza una agrupación “natural” de los datos.
- En cada etapa jerárquica el *cluster* padre a ser dividido se selecciona con base a un criterio que evalúa la calidad de las agrupaciones en función de la cohesión y separación de los *clusters*.

El esquema de la Figura 18 muestra el modo de operación del GBCM.

Figura 18. Esquema GBCM



Existen diferentes trabajos de investigación en los que se reseñan algunos criterios para seleccionar el *cluster* a dividir en etapas posteriores (índice de validación), cuando se aplica un método jerárquico divisivo [67] [68] [69]. Entre los cuales se resumen:

1. Cada *cluster* debe tener una división posterior, hasta alcanzar un criterio de parada definido. Es un criterio muy básico que ignora por completo la calidad de los *clusters* formados.
2. El *cluster* con mayor número de elementos es el que pasa a la etapa de división. Es el criterio con el que inicialmente se presentó el método B-KMA. Es un criterio simple, tiene la ventaja de formar ramas del dendograma equilibradas (de tamaños proporcionales). La desventaja radica en que no basa su elección en la calidad de los *clusters* formados.
3. El *cluster* con mayor varianza con respecto a sus Centroides pasa a la etapa de división. Es el más utilizado, ya que para su elección se basa en la calidad de los *clusters* formados, principalmente en las distancias *inter-cluster* e *intra-cluster*.

Para el presente trabajo se tiene en cuenta el criterio 3 para seleccionar el *cluster* a dividir en cada etapa jerárquica. Como índice de validación se propone usar una variante del índice Xie-Beni generalizado:

$$\text{Ind. Valid} = \frac{J(\mathbf{X}; \mathbf{U}, \mathbf{V})}{(d_{prom})^2} \quad (19)$$

Donde,

$d_{prom} = \text{prom}_{i,c} d^2(\mathbf{V}_i, \mathbf{V}_c)$, es la distancia promedio entre Centroides del *cluster*.

Con el criterio de validación de la ecuación (19) se busca minimizar la función objetivo, maximizando la distancia entre centroides de los *clusters*.

La función objetivo a minimizar del GBCM es la misma de la ecuación (5), con m igual a 2.

$$J(\mathbf{X}; \mathbf{U}, \mathbf{V}) = \sum_{i=1}^N \sum_{j=1}^k (u_{ij})^2 d^2(\mathbf{x}_i, \mathbf{v}_j)$$

Como criterio de parada para el GBCM se propone:

$$k \geq 30$$

(20)

Donde,

k Corresponde al número total de *clusters* alcanzados.

El criterio de parada se elige a criterio experto del analista, en el cual se considera que con 30 agrupaciones de los datos ya se cuenta con suficiente diversidad en los *clusters* formados para tomar decisiones en el direccionamiento.

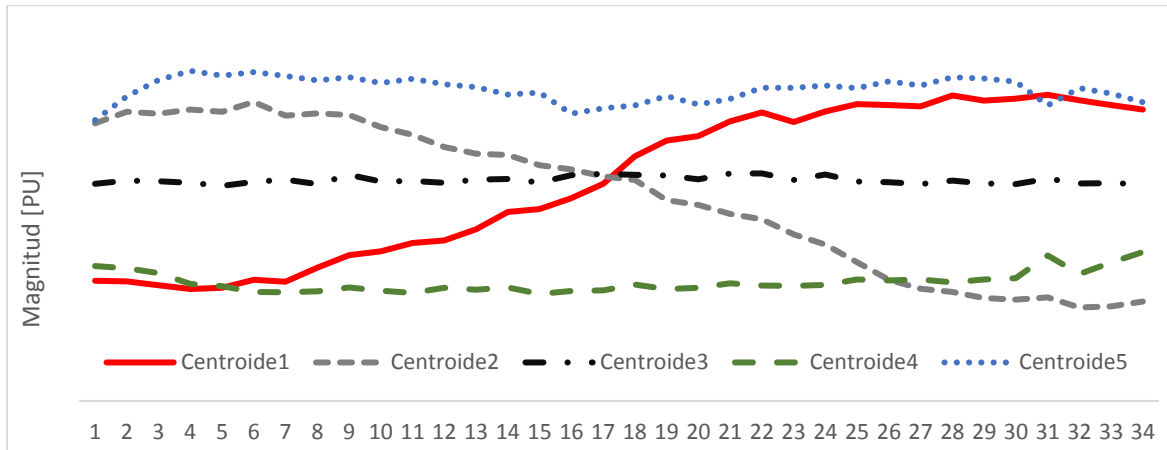
El método finaliza cuando se alcance el criterio de parada establecido, presentando como soluciones las obtenidas en las etapas terminales de las ramas del dendograma formado.

4. RESULTADOS Y CONCLUSIONES

4.1. AGRUPACIÓN CON AG

A las series de la Tabla 2 se les aplicó el método de agrupación basado en AG, para las cuales se obtuvieron los siguientes centroides asociados a los *clusters*:

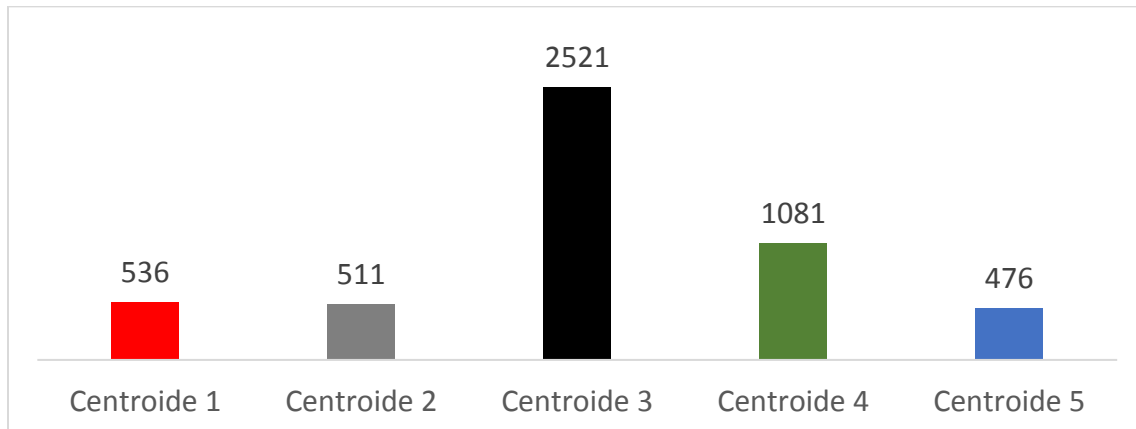
Figura 19. Centroides obtenidos aplicando AG



Se puede concluir que los datos se agruparon en $k=5$ *clusters*, a los cuales se les asigna una etiqueta lingüística de acuerdo a la tendencia observada en el centroide: magnitud alta (centroide 5), magnitud baja (centroide 4), magnitud media (centroide 3), de tendencia creciente (centroide 1) y tendencia decreciente (centroide 2).

El número de series agrupadas en cada *cluster* se observan en la Figura 20.

Figura 20. Distribución de series en la clasificación AG



El *cluster 3* es el más poblado con 2521 objetos, se asocia este fenómeno a que es el grupo de magnitud media, por lo cual se considera normal este comportamiento.

El resumen de los resultados obtenidos con la ejecución del AG se encuentra en la Tabla 3.

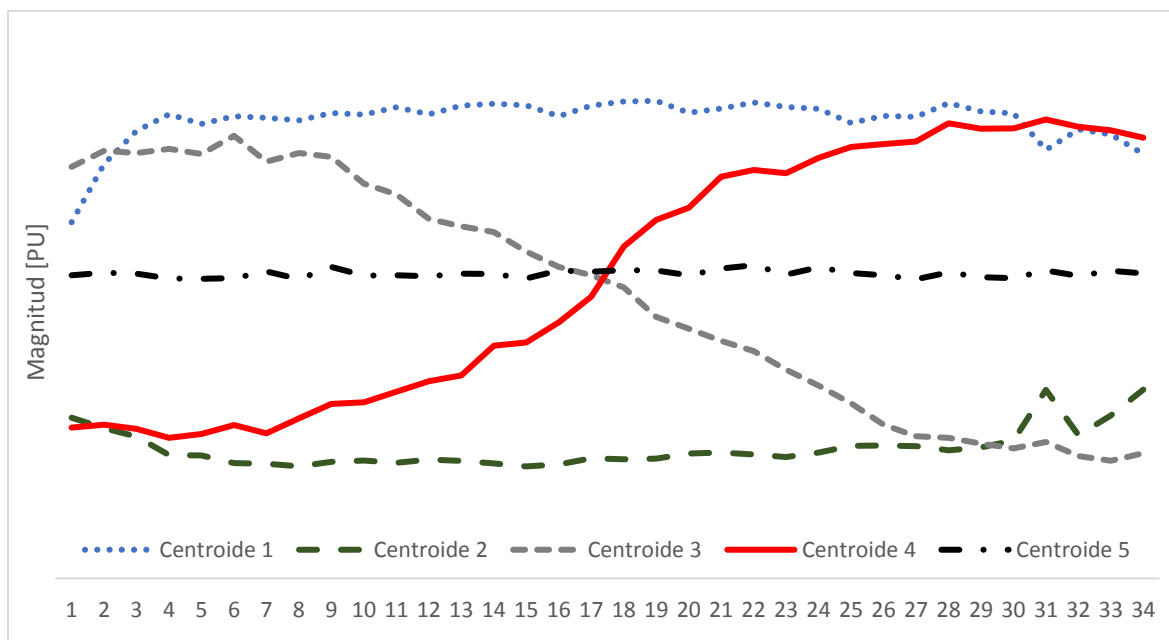
Tabla 3. Resumen ejecución AG

Número total de iteraciones	100
Tiempo de ejecución [segundos]	570,88
Función Objetivo alcanzada	1,8587
Número total de <i>clusters</i>	5

4.2. AGRUPACIÓN CON FCM-AG

Se aplicó la estrategia híbrida FCM-AG al conjunto de series de tiempo caracterizadas en la Tabla 2. En la Figura 21 se muestran los centroides para los *clusters* formados:

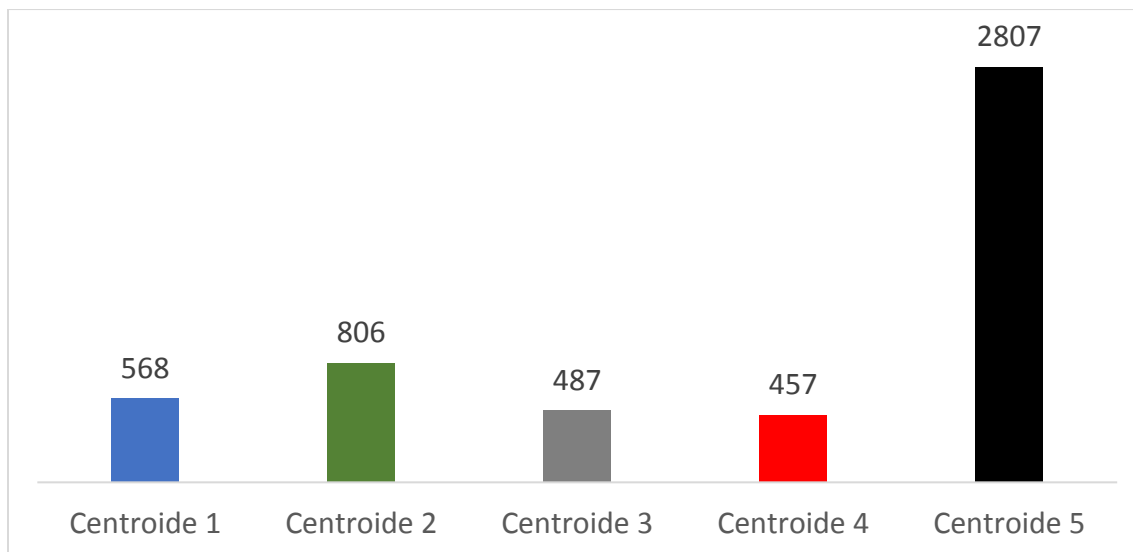
Figura 21. Centroides obtenidos aplicando FCM-AG



Para fines de agrupación de las series, se asume que estas se agrupan en el *cluster* con el cual tiene mayor porcentaje de pertenencia. De la Figura 21 se concluye que los datos fueron agrupados en 5 *clusters*, igual como sucedió en el AG.

El número de series que se agrupan en cada *cluster* se observan en la Figura 22.

Figura 22. Distribución de series en la clasificación FCM-AG



Al igual que en el AG, el *cluster* más poblado corresponde al de magnitud media, con 2807 objetos.

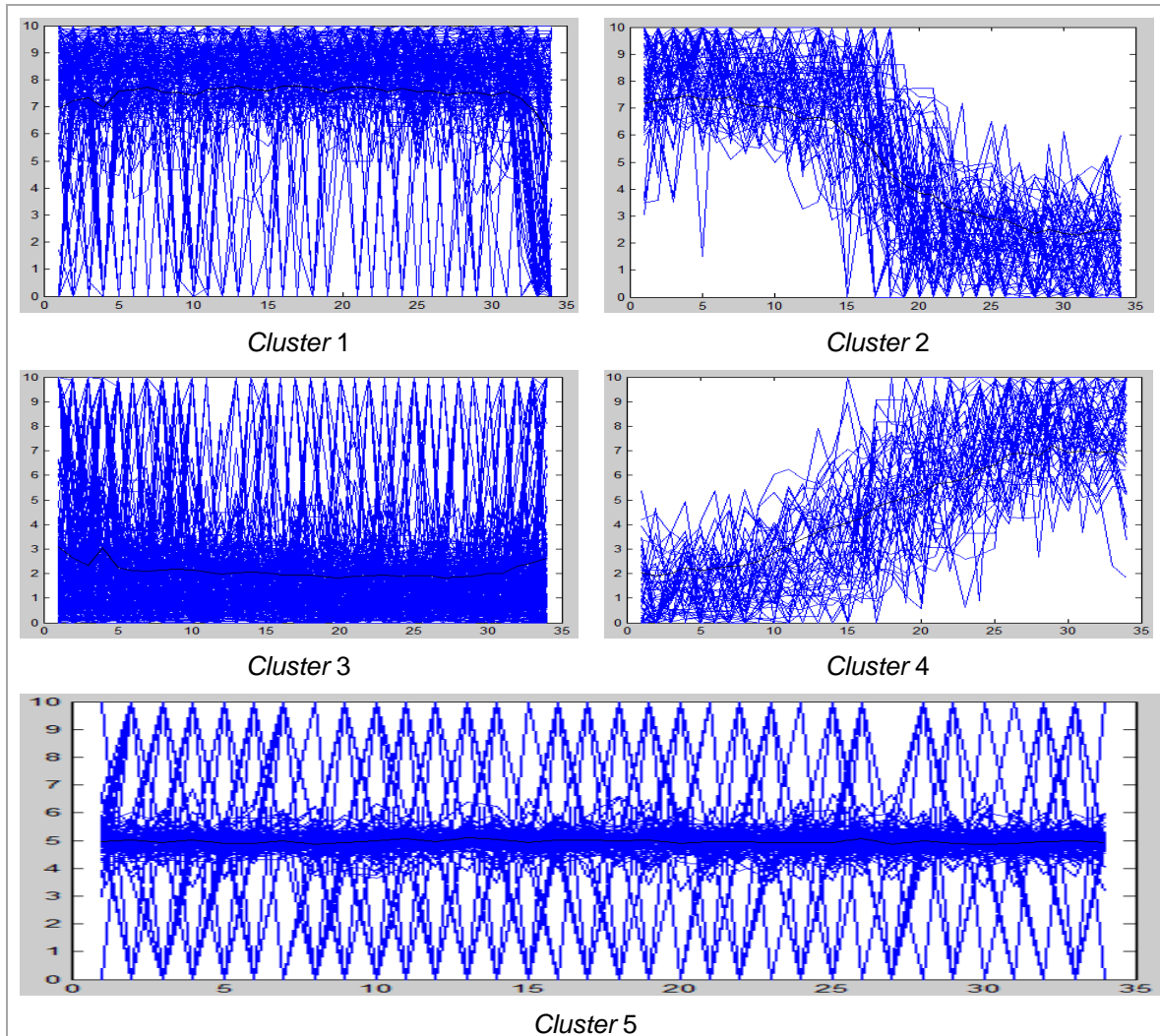
En la Tabla 4 se resumen los resultados obtenidos con el FCM-AG.

Tabla 4. Resumen ejecución FCM-AG

Número total de iteraciones	100
Tiempo de ejecución [segundos]	3994,43
Función Objetivo alcanzada	1,8733
Número total de <i>clusters</i>	5

Un resultado de gran relevancia para este trabajo es el porcentaje de pertenencia de las series a cada uno de los centroides. Este porcentaje de pertenencia se obtiene de aplicar el FCM como agrupador de los datos en la estrategia híbrida FCM-AG. En la Figura 23 se observan los *clusters* con las series que registran un valor de pertenencia superior al 50%.

Figura 23. *Cluster* obtenidos por FCM-AG



Aunque el porcentaje de pertenencia de las series mostradas supera el 50%, se observan picos que difieren un poco del centroide. Sin embargo, alrededor de éste se observa una zona densa que valida la correcta agrupación.

Las desviaciones en la tendencia entre las series y los centroides de un mismo *cluster* se pueden presentar por:

1. Series con alto contenido de ruido, el cual no logró ser normalizado en la etapa de pre-procesamiento y normalización de los datos.
2. Series altamente dinámicas, las cuales presentan características de conmutación por periodos de tiempo.

Para ilustrar el efecto de conmutación entre *clusters*, se presentan los siguientes tres casos:

Figura 24. Conmutación series de tiempo



Durante el periodo de análisis de la Figura 24 se observa que para los tres casos presentados, las series de tiempo asociadas a los transformadores conmutan entre dos centroides diferentes. La siguiente tabla refuerza este concepto:

Tabla 5. Porcentaje de pertenencia transformadores con conmutación

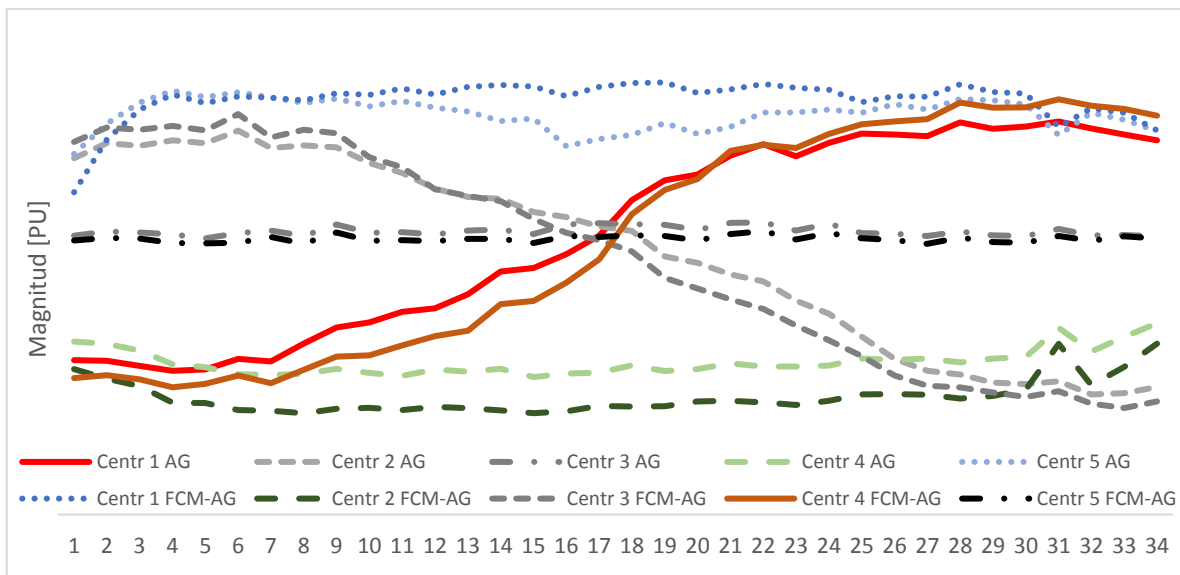
	% Difuso de pertenencia				
	Cluster 1	Cluster 2	Cluster 3	Cluster 4	Cluster 5
Transformador 3	9%	14%	11%	33%*	32%*
Transformador 4	8%	13%	37%*	9%	32%*
Transformador 5	31%*	7%	17%	14%	32%*

En la Tabla 5 se resaltan los valores de pertenencia más altos para cada transformador del ejemplo*. Estos valores coinciden con los *clusters* entre los cuales la serie asociada al balance de pérdidas del transformador presentó conmutación. Por ejemplo, la serie asociada a los balances de pérdidas del transformador 3 presentó conmutación entre los *clusters* 4 y 5. Por lo cual, dadas las características de conmutación que puedan presentar las series de tiempo a lo largo de un periodo de observación, no sería muy acertado clasificar la serie en un solo *cluster*, ya que iría en contra del comportamiento natural de la misma. Lo anterior afirma la validez de la aplicación del FCM como estrategia para encontrar agrupaciones naturales de los datos.

4.3. COMPARACIÓN DE LAS METODOLOGÍAS AG Y FCM-AG



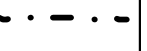







Comparando los *clusters* obtenidos por cada una de las metodologías se observa que son relativamente semejantes. La Figura 25 muestra los centroides propuestos por cada una.

Figura 25. Centroides generados por las metodologías AG y FCM-AG



El número de series que se agrupan en los *clusters* por cada metodología, se muestran en la Tabla 6.

Tabla 6. Distribución de agrupación de las series con AG y FCM-AG

AG \ FCM - AG						Total FCM-AG
	36	9	110		413	568
				806		806
		453	1	29	4	487
	432			22	3	457
	68	49	2410	224	56	2807
Total AG	536	511	2521	1081	476	

En la Tabla 6 se resaltan las series que se clasifican en las dos metodologías en el mismo *cluster*. A simple vista se observa que en los dos métodos un porcentaje alto de las series se clasifican en *clusters* con igual tendencia, sin embargo para comparar los resultados se utilizan los índices *Precision* (también llamado *Positive Predictive Value*) y *Recall* (también llamado *Sensitivity*) y la función *F_Measure* [22 - 34].

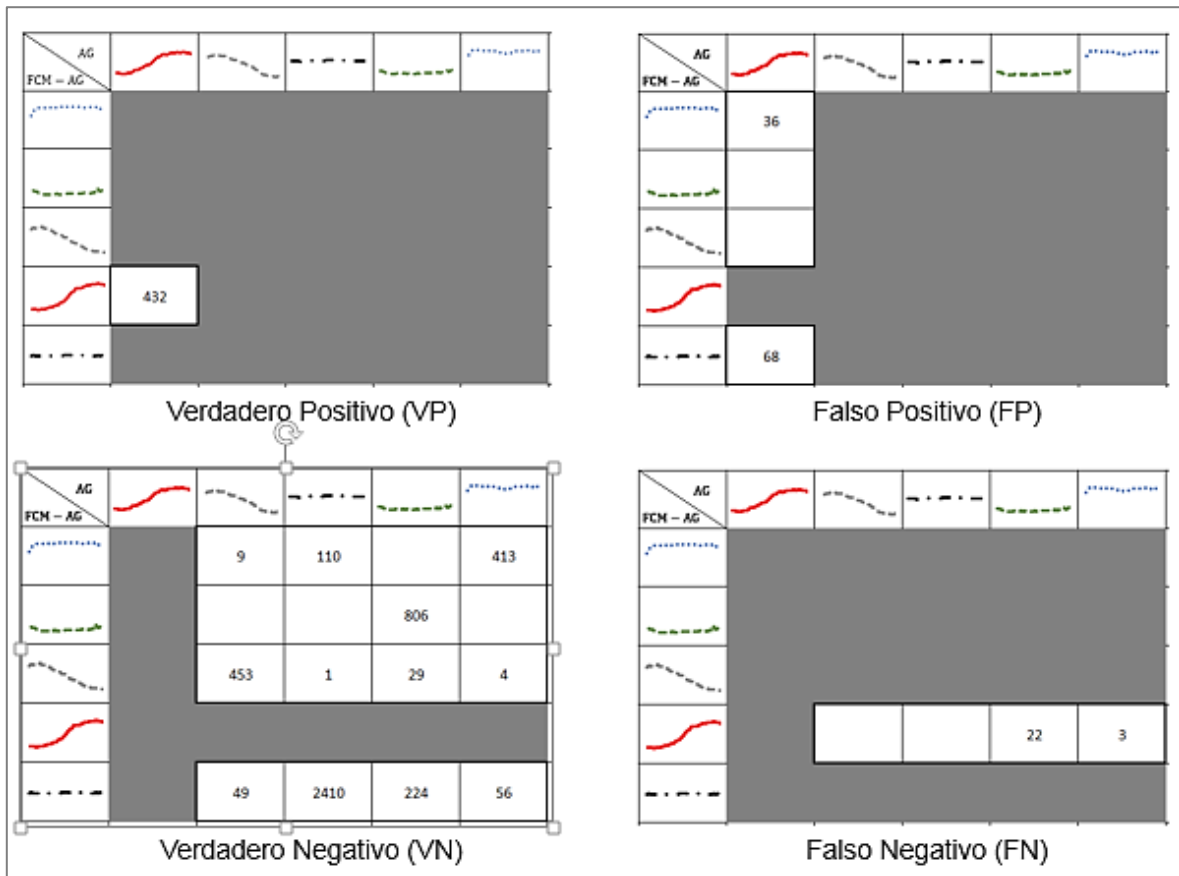
Tomando como referencia la clasificación realizada por el FCM-AG, dadas sus características de agrupación difusa que permite la agrupación natural de las series, se plantean los siguientes conceptos:

- Verdadero Positivo (VP): Corresponde a las series agrupadas por las dos metodologías en el mismo *cluster*. Corresponde a aquellas que se resaltan en la Tabla 6.
- Falso Positivo (FP): Corresponde a las series que se agrupan por el AG en el *cluster* de análisis, pero en el FCM-AG se agrupan en *clusters* diferentes.
- Verdadero Negativo (VN): Corresponde a las series agrupadas por las dos metodologías en un *cluster* diferente al analizado.

- Falso Negativo (FN): Corresponde a las series que se agrupan por el FCM-AG en el *cluster* de análisis, pero en el AG se agrupan en otro diferente.

Para ilustrar los conceptos anteriores se presenta la Figura 26 donde se aplican para el caso del *cluster* con tendencia creciente. En blanco aparece el valor seleccionado.

Figura 26. Ejemplo conceptos de agrupación



A partir de los anteriores conceptos, se formulan los índices *Precisión* y *Recall*.

$$Precision[\%] = \frac{VP}{VP + FP} \quad (21)$$

$$Recall[\%] = \frac{VP}{VP + FN} \quad (22)$$

El mejor de los escenarios para validar los métodos empleados es aquel en el que los dos índices anteriores se acerquen al 100%, lo cual se denomina *Utilidad Teórica*. Con el fin de evaluar que tan alejados se encuentran de la *Utilidad Teórica* se combinan ambos en una medida armónica denominada *F_Measure* [22].

$$F_Measure[\%] = \frac{2 \times Recall \times Precision}{Recall + Precision} \quad (23)$$

La Tabla 7 muestra los resultados obtenidos:

Tabla 7. Índices de comparación FCM-AG y AG

Clasificación	VP	FP	VN	FN	Precision	Recall	F_Measure
Magnitud Alta	413	63	4494	155	87%	73%	79%
Magnitud Baja	806	275	4044	0	75%	100%	85%
Magnitud Media	2410	111	2207	397	96%	86%	90%
Tendencia Decreciente	453	58	4580	34	89%	93%	91%
Tendencia Creciente	432	104	4564	25	81%	95%	87%
Valor promedio					85%	89%	87%

El valor promedio de todos los índices se encuentra por encima del 80%, lo cual indica que las dos metodologías comparadas se validan una a la otra, presentando resultados cercanos entre sí.

El resumen de los resultados obtenidos con la ejecución de las dos metodologías se encuentra en la Tabla 8.

Tabla 8. Comparación ejecución FCM-AG y AG

	FCM-AG	AG
Número total de iteraciones	100	100
Tiempo de ejecución [segundos]	3994,43	570,88
Función Objetivo alcanzada	1,8733	1,8587
Número total de <i>clusters</i>	5	5

Los métodos comparados agrupan los datos en 5 *clusters*, con un valor promedio de la función $F_Measure$ del 87%, lo que precisa una utilidad teórica aceptable de los resultados obtenidos por las dos metodologías. Aunque el método AG es más rápido en su ejecución respecto al FCM-AG, el dato no es relevante para este análisis ya que no se ejecuta en tiempo real. Incluso, para una empresa el periodo de aplicación de la metodología es semestral, ya que como los balances se generan mensualmente, seis meses son justos para descubrir con certeza si los transformadores antes direccionados han reaccionado y permite obtener una tendencia para los transformadores nuevos que entran en el análisis.

Si bien la función objetivo alcanzada en el AG es ligeramente menor a la obtenida por el FCM-AG, es más relevante para el análisis el valor agregado que da a la agrupación de las series el porcentaje de pertenencia a los *clusters* obtenido de aplicar el FCM-AG. Por lo anterior se considera de mayor peso en este trabajo la agrupación obtenida a partir del FCM-AG, ya que a partir del grado de pertenencia se pueden establecer estrategias para priorizar el direccionamiento en terreno y se elimina la incertidumbre de las series que se agrupan simultáneamente en diferentes grupos.

4.4. AGRUPACIÓN CON GBCM

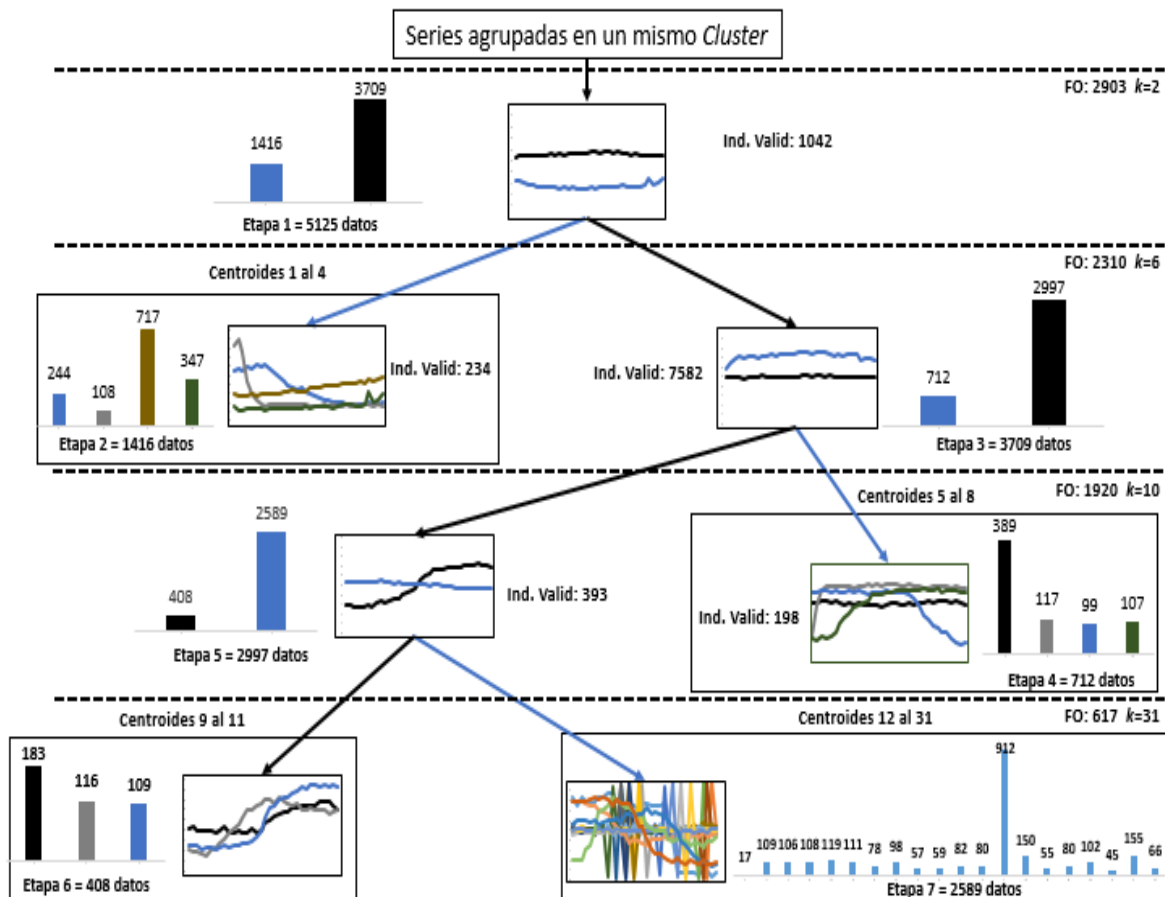
Las series a agrupar mediante el GBCM son las mismas que se agruparon con las metodologías anteriores.

El GBCM es un método de agrupación jerárquico divisivo, cuyo proceso de agrupamiento se ilustra en forma de dendograma en la Figura 27. A continuación se describe el proceso seguido:

- El proceso inicia con todas las series agrupadas en un mismo *cluster*.
- Para llegar al agrupamiento obtenido en la etapa 1, las series son divididas aplicando el FCM-AG. Para este caso el rango de variación del individuo generado en el AG se fija entre 2 y 30, relajando en una unidad la restricción de mínimos (*black holes*). Como resultado las series se agrupan en dos *clusters*, los cuales de acuerdo a la forma de sus centroides se pueden explicar como de magnitud media y mínima, respectivamente.
- Una vez se verifica el criterio de parada establecido en la ecuación (18), los *clusters* que integran la Etapa 1 pasan de manera individual al proceso de agrupación. Es así como de la etapa padre (Etapa 1) surgen dos nuevas etapas (Etapa 2 y Etapa 3). La Etapa 2 se forma del proceso de agrupamiento de las series contenidas en el *cluster* 1 de la Etapa 1. La Etapa 3 se forma agrupando las series contenidas en el *cluster* 2 de la Etapa 1. Hasta este punto se han obtenido $k=6$ *clusters*.

- Se verifica nuevamente el criterio de parada, si no se cumple se ejecuta el siguiente nivel jerárquico de agrupación. La etapa a agrupar corresponde a aquella con mayor índice de validación (*Ind. Valid*), es así como la Etapa 3 es la seleccionada.
- El proceso se repite de manera iterativa hasta alcanzar el criterio de parada establecido.

Figura 27. Dendograma obtenido de aplicar el GBCM



Para este caso particular el proceso finaliza con las series agrupadas en $k=31$ clusters. Los centroides se obtienen de los ramales terminales de la Figura 27 que corresponden a las etapas 2, 4, 6 y 7.

El GBCM cumple las características propias de un método jerárquico divisivo, en el cual se puede aplicar el proceso de división consecutivamente hasta que cada serie forme su propio cluster.

Algunas de las bondades encontradas al aplicar el GBCM son las siguientes:

- Al ser un método de tipo jerárquico se mitiga la restricción de escalabilidad del FCM de trabajar con grandes bases de datos. A medida que se van generando etapas las bases se hacen más pequeñas, por lo cual mejora la calidad en la agrupación.
- Al trabajar con un índice de validación de agrupación basado en la correlación (Pearson), se garantiza una agrupación de las series por forma, lo cual se comprueba con la tendencia que siguen los centroides de los *clusters* obtenidos.
- Al operar de manera colaborativa con el AG, se explora ampliamente el espacio de búsqueda, permitiendo que en cada etapa jerárquica el resultado obtenido se acerque al óptimo global del problema.
- La ventaja del GBCM con respecto al B-KMA es que las condiciones iniciales de operación del método de agrupación (en este caso el FCM) son dadas por el AG, por lo cual no se restringe a una clasificación de dos *cluster* por etapa jerárquica, sino que permite encontrar en cada etapa los *clusters* naturales a la forma de las series.
- Utilizando el enfoque difuso jerárquico se mejora el rendimiento computacional, ya que la agrupación difusa permite conservar la estructura natural de las series de tiempo haciendo que se requieran menos iteraciones para la convergencia en la agrupación.

Se resalta la bondad que tiene el GBCM en centrar la clasificación basado en los detalles de la forma de las series de tiempo. Esta afirmación se puede ver reflejada en el paso de la etapa 5 a la etapa 6. En este caso el *cluster* 1 de la etapa 5 tiene una tendencia creciente, sin embargo al ser dividido en la etapa 6, se forman tres *clusters* donde los centroides conservan la misma tendencia creciente, pero centrándose más en el detalle de la forma, como el momento a partir del cual las pérdidas comienzan a incrementar.

Con el fin de validar las agrupaciones, se hace una comparación de los *clusters* obtenidos con el GBCM y el FCM-AG.

4.5. COMPARACIÓN RESULTADOS FCM-AG Y GBCM

Como ya se mencionó anteriormente, el FCM-AG es un método particional que agrupa los datos en un solo paso jerárquico, centrándose en la generalidad de la tendencia de las series como criterio de agrupación. Por otro lado, el GBCM opera como un método jerárquico divisivo, el cual en cada etapa del proceso ejecuta el FCM-AG para la re-agrupación del *cluster* padre en análisis. Como criterio de

agrupamiento, el GBCM se centra en detalles particulares de la forma de las series analizadas. Ambos métodos conservan la ventaja que ofrece el enfoque difuso en el agrupamiento de series reales, cuyo comportamiento es difícil encasillarlo en un único *cluster*. Igualmente, en ambos métodos se aprovecha la característica del AG de explorar ampliamente el espacio de solución con el fin de encontrar el número de *clusters* k más adecuado para la agrupación. Operan bajo la filosofía de los métodos No supervisados, por lo cual presentan la ventaja de que no requieren de etiquetas previas en los datos, las cuales en problemas reales no son muy comunes.

Con el fin de establecer una relación entre las dos metodologías, se plantea una tabla de equivalencia entre *clusters* encontrados por cada una.











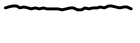

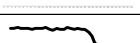
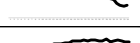



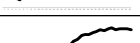
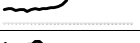


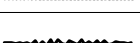

Tabla 9. Equivalente de agrupación entre el GBCM y el FCM-AG

Clasificación	Descripción
	<i>clusters</i> de tendencia decreciente
	<i>clusters</i> Magnitud baja
	<i>clusters</i> Magnitud alta
	<i>clusters</i> de tendencia creciente
	<i>clusters</i> Magnitud media

La Tabla 9 permite realizar una comparación de *clusters* entre las dos metodologías utilizadas con el fin de encontrar puntos en común que permitan validar los resultados.

En la Tabla 10 se presenta un resumen de los *clusters* encontrado por las dos metodologías, y en esta se implementa la codificación descrita en la Tabla 9.

Tabla 10. Resumen de *clusters* obtenidos por FCM-AG y GBCM

FCM-AG		Centroide 1	Centroide 2	Centroide 3	Centroide 4	Centroide 5
GBCM						
Centroide 1			51	170		23
Centroide 2			96	7		5
Centroide 3			312		107	298
Centroide 4			347			
Centroide 5		329			11	49
Centroide 6		117				
Centroide 7		61		35		3
Centroide 8		61			44	2
Centroide 9					99	84
Centroide 10					87	29
Centroide 11					109	
Centroide 12				17		
Centroide 13 al 23 y 26 al 28				15		1229
Centroide 24				4		908
Centroide 25				68		82
Centroide 29				14		31
Centroide 30				91		64
Centroide 31				66		

Igual como se estableció en la sección 4.3, se plantean los índices *Precision*, *Recall* y la función *F_Measure* con el fin de validar los resultados obtenidos. En este caso se toma como referencia la clasificación realizada por el GBCM, considerando que la agrupación obtenida por este método se basa en la particularidad de la tendencia de las series.

De acuerdo a lo anterior, se redefinen los siguientes conceptos:

- Verdadero Positivo (VP): Corresponde a las series agrupadas por las dos metodologías en el mismo *cluster*.
- Falso Positivo (FP): Corresponde a las series que se agrupan por el FCM-AG en el *cluster* de análisis, pero por el GBCM se agrupan en *clusters* diferentes.
- Verdadero Negativo (VN): Corresponde a las series agrupadas por las dos metodologías en un *cluster* diferente al analizado.
- Falso Negativo (FN): Corresponde a las series que se agrupan por el GBCM en el *cluster* de análisis, pero en el AGBCM se agrupan en uno diferente.

La Tabla 11 muestra los resultados obtenidos:

Tabla 11. Índices de comparación FCM-AG y GBCM

Clasificación	VP	FP	VN	FN	<i>Precision</i>	<i>Recall</i>	<i>F_Measure</i>
	468	19	4222	416	96%	53%	68%
	659	147	3914	405	82%	62%	70%
	329	239	4497	60	58%	85%	69%
	339	118	4492	176	74%	66%	70%
	2137	670	2299	19	76%	99%	86%
Valor promedio					77%	77%	73%

El *F_Measure* promedio de todos los *clusters* es del 73%, por lo cual se puede concluir en términos generales que los dos métodos de agrupamiento presentan un aceptable índice de *Utilidad Teórica* y que aunque cada uno presente una lógica de agrupación diferente comparten una coherencia en la agrupación.

Con los resultados obtenidos (*Precisión* promedio 77%, *Recall* promedio 73%) no es posible afirmar que un método sea mejor que el otro, sin embargo, se validan las

características de agrupación de cada uno. El GBCM permite conocer con más detalle los datos analizados, lo que facilita tomar decisiones basadas en particularidades. El FCM-AG puede ser aplicado cuando el objetivo del análisis es conocer de manera general la estructura de los datos.

4.6. ANÁLISIS DE TRANSFORMADORES SOSPECHOSOS

Con el fin de identificar los transformadores con comportamiento “sospechoso” que adviertan la existencia de pérdidas no técnicas, se hace una revisión de las series agrupadas en los *clusters* de tendencia creciente mediante las metodologías (FCM-AG y GBCM).

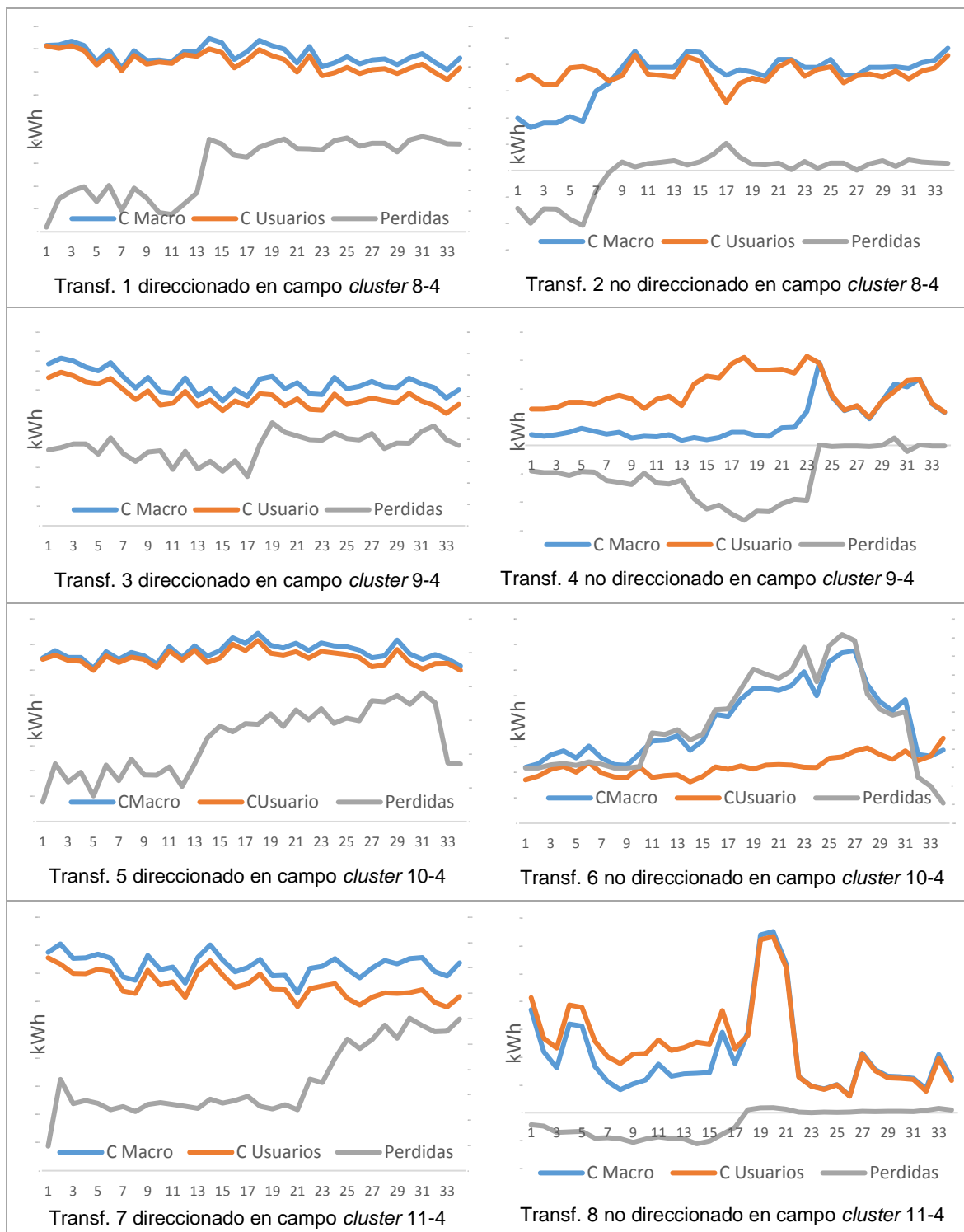
Se revisaron las series agrupadas en los *clusters* 8-4 (8 GBCM y 4 FCM-AG), 9-4 (9 GBCM y 4 FCM-AG), 10-4 (10 GBCM y 4 FCM-AG) y 11-4 (11 GBCM y 4 FCM-AG) con el fin de identificar algún patrón que permita concluir al respecto. Para el análisis se involucró el comportamiento histórico del consumo de los macromedidores y los usuarios. Los resultados se resumen en la Tabla 12.

Tabla 12. Análisis *clusters* 8-4, 9-4, 10-4 y 11-4

	Direccionar	No Direccionar
<i>Cluster</i> 8-4	13	31
<i>Cluster</i> 9-4	37	62
<i>Cluster</i> 10-4	32	55
<i>Cluster</i> 11-4	59	50
Total	141	198

De los *clusters* analizados, se direccionaron para revisión en campo 141 transformadores, los cuales, complementando con el análisis de tendencia del consumo del macromedidor y los usuarios tienen un comportamiento de pérdidas “sospechoso”. A la fecha de elaboración de este documento, de todos los transformadores direccionados 31 ya fueron inspeccionados en campo con una efectividad del 100%. Los problemas identificados son: usuarios ilegales, fraudes, anomalías, asentamientos humanos, entre otros. En la Figura 28 se muestran casos particulares del comportamiento de los transformadores analizados.

Figura 28. Transformadores analizados por *clusters* 8-4, 9-4, 10-4 y 11-4



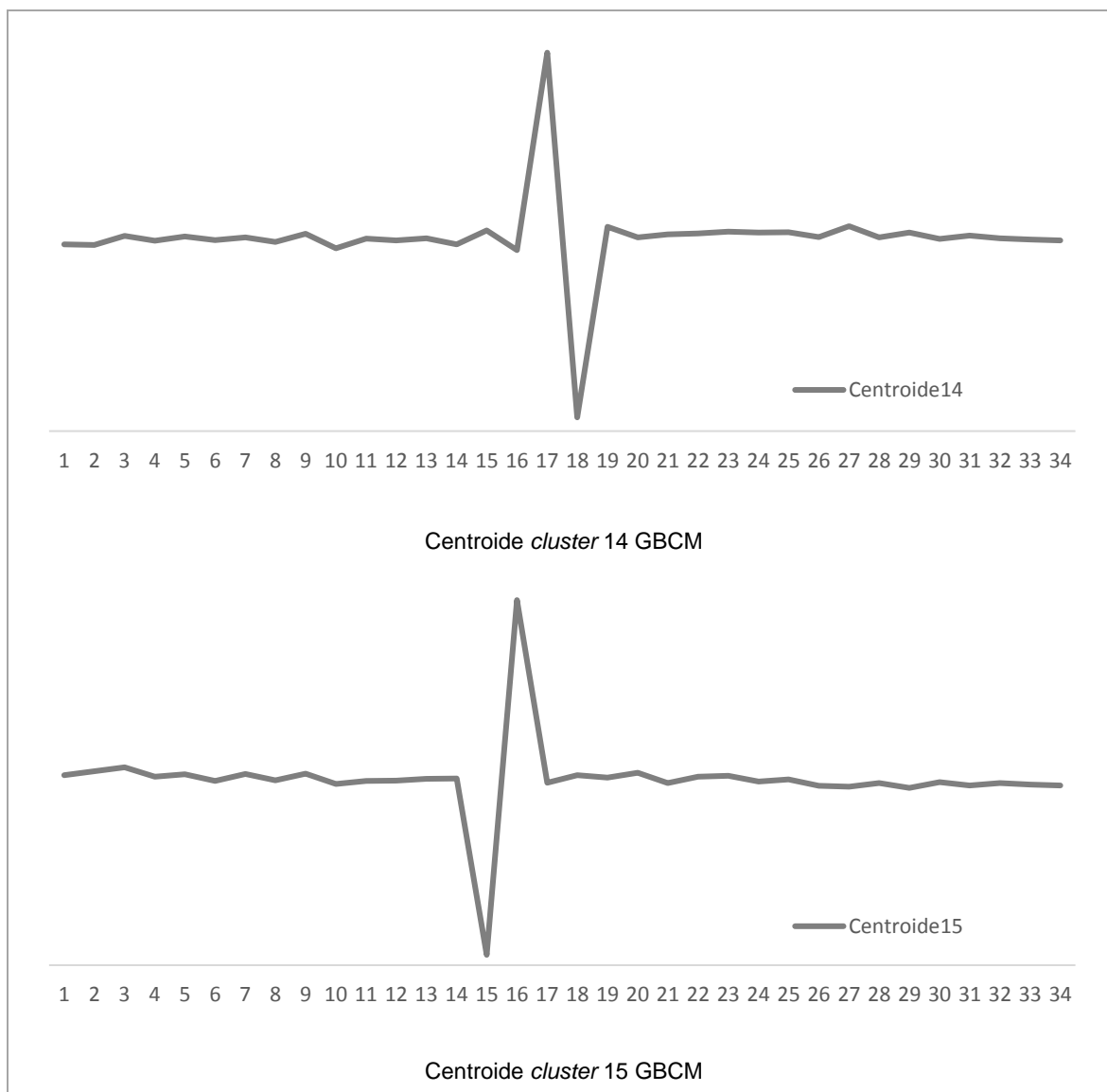
Según el comportamiento de los transformadores de la Figura 28, se tienen las siguientes observaciones:

- Los transformadores clasificados en el *cluster* 8-4 tienen la particularidad que las pérdidas incrementan de manera significativa y casi instantánea en los primeros meses de la serie. El transformador 2 de la Figura 26 no fue direccionado a terreno porque el incremento se dio por un paso de pérdidas negativas a positivas (cercanas a cero) dada una reparación del macromedidor.
- Las pérdidas de los transformadores clasificados en el *cluster* 9-4 también tienen un incremento, el cual se localiza hacia la mitad de la serie, sin embargo muestra una leve recuperación hacia el final de la misma. Al igual que en el caso anterior, el transformador 4 no se direcciona ya que el incremento en pérdidas corresponde a una reparación del macromedidor.
- Para el *cluster* 10-4 las pérdidas se incrementan considerablemente, pero presentan una recuperación importante hacia el final de la serie. El transformador 6 no es direccionado a campo ya que la recuperación en pérdidas lo lleva a un valor cercano a cero.
- Los transformadores agrupados en el *cluster* 11-4 presentan un incremento en pérdidas significativo, este incremento no se presenta de manera instantánea, sino de manera gradual en el tiempo. Al igual que los casos anteriores, el transformador 8 no se direcciona ya que corresponde a una reparación de macromedidor.

Como se observa en la Figura 28, en todos los *clusters* de tendencia creciente se presentan transformadores que de acuerdo al análisis previo ameritan o no un direccionamiento a terreno para inspección de pérdidas. De aquí la importancia de correlacionar en un análisis de pérdidas el consumo de los usuarios, macromedidores y las pérdidas totales. A partir del análisis de estas tres componentes el analista de pérdidas de la empresa tiene un panorama más amplio para seleccionar de manera más acertada los transformadores “sospechosos” de pérdidas.

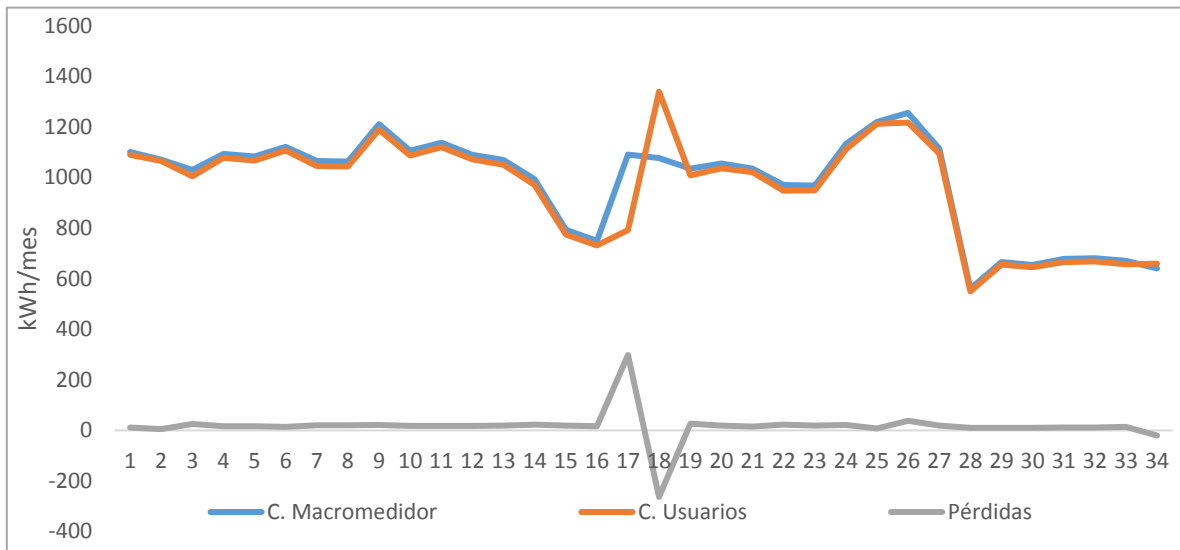
El enfoque de este trabajo fue el de encontrar transformadores que por su tendencia histórica en las pérdidas se pueden etiquetar como “sospechosos” de fraude y por lo cual ameritan un direccionamiento para ser inspeccionados en terreno. Sin embargo, este análisis puede ser más amplio y abarcar otro tipo de direccionamiento en terreno como el caso de macromedidores con fallas, irregularidades en tomas de lectura y errores de facturación. Estos se obtienen del análisis de los *clusters* restantes. Para dar más claridad acerca del tema se muestran dos casos particulares obtenidos a partir de los centroides de los *clusters* 14 y 15 del GBCM.

Figura 29. Transformadores clasificados en *clusters* 14 y 15 GBCM



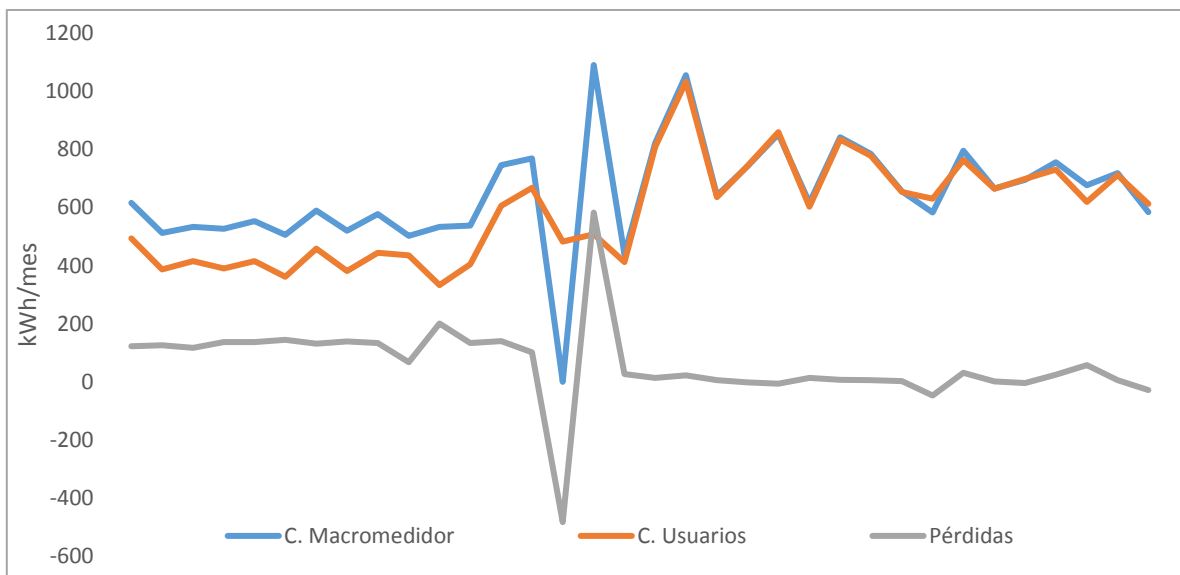
El centroide del *cluster* 14 presenta este comportamiento por usuarios que ingresan al sistema de distribución en el mes 17, sin embargo no fueron actualizados en el sistema comercial hasta el mes siguiente. Por lo anterior, en el mes 17 el macromedidor registra el incremento de las pérdidas, pero para el mes 18 los usuarios nuevos facturan dos meses acumulados, por lo cual el balance es negativo. Para el mes 19 los balances de energía regresan a la tendencia que traían. En la Figura 30 se revisa un caso particular, involucrando en el análisis el consumo del macromedidor y los usuarios.

Figura 30. Transformador 1 analizado por error en la facturación



El centroide presentado en el *cluster* 15 se da en los casos para los cuales no se toma lectura en el mes 15 para el macromedidor, por lo cual el balance es negativo durante ese mes. Para el mes siguiente (mes 16) se factura el macromedidor con una lectura acumulada, por lo cual el balance de energía crece. Para los meses siguientes el balance regresa a la tendencia que traía. En la Figura 31 se muestra este efecto sobre un caso particular para el cual no se normalizan los datos y se involucra en el análisis la facturación histórica del macromedidor y los usuarios.

Figura 31. Transformador 2 analizado por error en la facturación



Además de lo particular de la fluctuación en la facturación del macromedidor entre los meses 15 y 16, en la Figura 31 se observa que después del fenómeno transitorio se regulan las pérdidas y bajan a un valor cercano a cero, lo cual sucede por el incremento en el promedio de facturación de los usuarios. No se identifica en el sistema comercial de la empresa de distribución una explicación para este hecho, sin embargo puede ser una línea más de análisis a la que se le puede dar un enfoque a partir de esta metodología.

5. TRABAJOS FUTUROS

Realizar un análisis de la base de datos trabajada con el fin de definir la pertinencia de normalizar las series para el análisis de *cluster* realizado, o en su defecto, para determinar si la normalización aplicada es la más apta con el fin de descubrir tendencias que permitan establecer un patrón de comportamiento de los transformadores con pérdidas.

En la etapa de pre-procesamiento aplicar un método de filtrado por serie, con el fin de obtener su forma básica sin considerar el ruido. Un ejemplo de estos es el “filtro de la ventana de Gauss” utilizado en [70]. De esta forma se pueden suavizar los picos y hundimientos de las series, permitiendo que para la agrupación de las series el método aplicado se concentre más en la tendencia en el tiempo de las mismas y no en los cambios repentinos que se puedan presentar.

Aplicar al proceso de agrupación un método que permita encontrar *clusters* en entornos ruidosos, de tal manera que identifique los puntos de ruido y los agrupe en un nuevo conjunto $k+1$. Un método de este tipo es presentado en [71].

Integrar en una misma metodología de análisis la tendencia de las series asociadas al consumo histórico de los macromedidores, los usuarios y las pérdidas del transformador. Esto con el fin de establecer reglas de asociación de la tendencia que permitan hacer direccionamientos de pérdidas con mayor precisión.

Complementar la metodología propuesta con otros atributos, que además de la tendencia del histórico de pérdidas, permita describir los transformadores por sus características de ubicación, consumo y uso. Estos atributos pueden ser: clase de servicio de los usuarios, estrato, ubicación física del transformador (urbano, rural, barrio, municipio), consumo promedio de transformadores vecinos (transformadores con iguales características), entre otros.

Aplicar enfoques clásicos de la minería de datos para la agrupación de series de tiempo, complementando con técnicas de extracción de características de dimensiones inferiores que puedan representar la serie. Algunas características útiles que se puedan extraer de las series temporales son: Funciones estimadas de autocorrelación (*estimated autocorrelation functions (ACFs)*), funciones de densidad espectral estimada (*estimated spectral density functions*) y estimación de la varianza wavelet.

6. BIBLIOGRAFIA

- [1] JIMÉNEZ, Raúl Alberto; SEREBRISKY, Tomás; MERCADO DÍAZ, Jorge Enrique. Electricidad perdida: Dimensionando las pérdidas de electricidad en los sistemas de transmisión y distribución en América Latina y el Caribe. IDB Monograph (Infrastructure and Environment Sector. Energy Division); IDB-MG-241, 2014.
- [2] NIZAR, A. H.; DONG, Z. Y.; WANG, Y. Power utility nontechnical loss analysis with extreme learning machine method. IEEE Transactions on Power Systems, 2008, vol. 23, no 3, p. 946-955.
- [3] CREG, Comisión de Regulación de Energía y Gas. Resolucion Creg 172 de 2011: Por la cual se establece la metodología para la implementación de los Planes de reducción de pérdidas no técnicas en los Sistemas de Distribución Local. Colombia. Diciembre 2011.
- [4] TREJOS RAMÍREZ, Rubén Darío. Metodología para la detección de pérdidas no técnicas en sistemas de distribución utilizando métodos de minería de datos. 2014. Tesis de Maestría. Pereira: Universidad Tecnológica de Pereira.
- [5] RAJABI, A.; LI, L.; ZHANG, J.; Zhu, J.; GHAVIDEL, S.; GHADI, M. J. A review on clustering of residential electricity customers and its applications. En Electrical Machines and Systems (ICEMS), 2017 20th International Conference on. IEEE, 2017. p. 1-6.
- [6] LI, Jiamin; LEWIS, Harold W. Fuzzy Clustering Algorithms—Review of the Applications. En Smart Cloud (SmartCloud), IEEE International Conference on. IEEE, 2016. p. 282-288.
- [7] HAN, Jiawei; PEI, Jian; KAMBER, Micheline. Data mining: concepts and techniques. Elsevier, 2011.
- [8] MACQUEEN, James. Some methods for classification and analysis of multivariate observations. En Proceedings of the fifth Berkeley symposium on mathematical statistics and probability. 1967. p. 281-297.
- [9] BEZDEK, James C.; EHRLICH, Robert; FULL, William. FCM: The fuzzy c-means clustering algorithm. Computers & Geosciences, 1984, vol. 10, no 2-3, p. 191-203.
- [10] XIE, Xuanli Lisa; BENI, Gerardo. A validity measure for fuzzy clustering. IEEE Transactions on pattern analysis and machine intelligence, 1991, vol. 13, no 8, p. 841-847.

- [11] BANDYOPADHYAY, Sanghamitra; MAULIK, Ujjwal. Genetic clustering for automatic evolution of clusters and application to image classification. *Pattern recognition*, 2002, vol. 35, no 6, p. 1197-1208.
- [12] DAS, Swagatam; ABRAHAM, Ajith; KONAR, Amit. Automatic clustering using an improved differential evolution algorithm. *IEEE Transactions on systems, man, and cybernetics-Part A: Systems and Humans*, 2008, vol. 38, no 1, p. 218-237.
- [13] MAULIK, Ujjwal; BANDYOPADHYAY, Sanghamitra. Fuzzy partitioning using a real-coded variable-length genetic algorithm for pixel classification. *IEEE Transactions on geoscience and remote sensing*, 2003, vol. 41, no 5, p. 1075-1081.
- [14] HOLLAND, John H. *Adaptation in natural and artificial systems. An introductory analysis with application to biology, control, and artificial intelligence*. Ann Arbor, MI: University of Michigan Press, 1975.
- [15] RAGHAVAN, Vijay V.; BIRCHARD, Kim. A clustering strategy based on a formalism of the reproductive process in natural systems. En *ACM SIGIR Forum*. ACM, 1979. p. 10-22.
- [16] BANDYOPADHYAY, Sanghamitra; MAULIK, Ujjwal. An evolutionary technique based on K-means algorithm for optimal clustering in RN. *Information Sciences*, 2002, vol. 146, no 1, p. 221-237.
- [17] PATERLINI, Sandra; MINERVA, Tommaso. Evolutionary approaches for cluster analysis. En *Soft Computing Applications*. Physica, Heidelberg, 2003. p. 165-176.
- [18] MAULIK, Ujjwal; BANDYOPADHYAY, Sanghamitra. Genetic algorithm-based clustering technique. *Pattern recognition*, 2000, vol. 33, no 9, p. 1455-1465.
- [19] CHIOU, Yu-Chiun; LAN, Lawrence W. Genetic clustering algorithms. *European journal of operational research*, 2001, vol. 135, no 2, p. 413-427.
- [20] KRISHNA, K.; MURTY, M. Narasimha. Genetic K-means algorithm. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, 1999, vol. 29, no 3, p. 433-439.
- [21] GARCÍA, Sergio López; LAYES, Luís Magdalena; PÉREZ, Juan Ramón Velasco. Algoritmo de agrupamiento genético borroso basado en el algoritmo de las C-medias borroso. 2001. Tesis Doctoral. Universidad Politécnica de Madrid.
- [22] STEINBACH, Michael; KARYPIS, George; KUMAR, Vipin. A comparison of Document Clustering Techniques. *Proceedings of World Text Mining Conference, KDD2000, Boston*.

- [23] BANERJEE, Shreya; CHOUDHARY, Ankit; PAL, Somnath. Empirical Evaluation of K-Means, Bisecting K-Means, Fuzzy C-Means and Genetic K-Means Clustering Algorithms. En Electrical and Computer Engineering (WIECON-ECE), 2015 IEEE International WIE Conference on. IEEE, 2015. p. 168-172.
- [24] MURUGESAN, Keerthiram; ZHANG, Jun. Hybrid bisect K-means clustering algorithm. En Business Computing and Global Informatization (BCGIN), 2011 International Conference on. IEEE, 2011. p. 216-219.
- [25] NAGI, J; MOHAMMAD, A. M.; YAP, K. S.; TIONG, S. K.; AHMED, S. K. Non-technical loss analysis for detection of electricity theft using support vector machines. En Power and Energy Conference, 2008. PECon 2008. IEEE 2nd International. IEEE, 2008. p. 907-912.
- [26] IRINEL-SORIN ILIE; GIANFRANCO C. Exploiting Support Vector Clustering Techniques for electrical Load Profiling. 2008. 2ND International conference on modern power systems. Romania. Pp. 12-14.
- [27] NIZAR, Anisah H.; DONG, Zhao Yang; ZHANG, Pei. Detection rules for non-technical losses analysis in power utilities. En Power and Energy Society General Meeting-Conversion and Delivery of Electrical Energy in the 21st Century, 2008 IEEE. IEEE, 2008. p. 1-8.
- [28] CABRAL, J. E; PINTO, J. O. P.; GONTIJO, E. M.; REIS, J. Rough sets based fraud detection in electrical energy consumers. En WSEAS International Conference on MATHEMATICS AND COMPUTERS IN PHYSICS, Cancun, Mexico. 2004. p. 413-416
- [29] MONEDERO, I.; BISCARRI, F.; LEON, C.; GUERRERO, J. I.; BISCARRI, J.; MILLAN, R. New methods to detect non-technical losses on power utilities. Catalonia, 2009, vol. 27695, no 598, p. 2.15.
- [30] MOLIN, A. D.; PEREIRA, J. O; CARVALHO, A.M; SAUER, L; COLMAN, E. Fraud Detection in Electric Energy Using Differential Evolution. Federal University of Mato Grosso do Sul. Brasil.
- [31] NIZAR, A. H.; DONG, Z. Y.; JALALUDDIN, M.; RAFFLES, M. J. Load profiling method in detecting non-technical loss activities in a power utility. En Power and Energy Conference, 2006. PECon'06. IEEE International. IEEE, 2006. p. 82-87.
- [32] NIZAR, A. H.; DONG, Z. Y.; ZHAO, J. H.; ZHANG, P. A data mining based NTL analysis method. En Power Engineering Society General Meeting, 2007. IEEE. IEEE, 2007. p. 1-8.
- [33] NAGI, J.; YAP, K. S.; TIONG, S. K.; AHMED, S. K.; MOHAMMAD, A. M. Detection of abnormalities and electricity theft using genetic support vector

machines. En TENCON 2008-2008 IEEE Region 10 Conference. IEEE, 2008. p. 1-6.

[34] ANGELOS, E. W. S.; SAAVEDRA, O. R.; CORTÉS, O. A. C.; DE SOUZA, A. N. Detection and identification of abnormalities in customer consumptions in power distribution systems. IEEE Transactions on Power Delivery, 2011, vol. 26, no 4, p. 2436-2442.

[35] BABU, T. Vandana; MURTHY, T. Satyanarayana; SIVAIAH, B. Detecting unusual customer consumption profiles in power distribution systems—APSPDCL. En Computational Intelligence and Computing Research (ICCIC), 2013 IEEE International Conference on. IEEE, 2013. p. 1-5.

[36] YAP, K. S.; TIONG, S. K.; NAGI, J.; KOH, J. S.; NAGI, F. Comparison of supervised learning techniques for non-technical loss detection in power utility. International Review on Computers and Software (I. RE. CO. S.), 2012, vol. 7, no 2, p. 1828-6003.

[37] Ley 142 de 1994. Por la cual se establece el régimen de los servicios públicos domiciliarios y se dictan otras disposiciones. Colombia. 1994.

[38] RAJARAMAN, Anand; ULLMAN, Jeffrey David. Mining of massive datasets. Cambridge university press, 2012.

[39] ELDER, John. Handbook of statistical analysis and data mining applications. Academic Press, 2009.

[40] RIGAUX, Philippe; SCHOLL, Michel; VOISARD, Agnes. Spatial databases: with application to GIS. Morgan Kaufmann, 2001.

[41] DE LA FUENTE, S. Análisis conglomerados. Facultad de Ciencias Económicas y Empresariales, Universidad Autónoma de Madrid (UAM). Madrid, España, 2011.

[42] VILLAGRA, Andrea; LEGUIZAMÓN, Guillermo. Metaheurísticas aplicadas a Clustering. Departamento de informática. Universidad Nacional de San Luis. Argentina.

[43] HALKIDI, Maria; BATISTAKIS, Yannis; VAZIRGIANNIS, Michalis. On clustering validation techniques. Journal of intelligent information systems, 2001, vol. 17, no 2, p. 107-145.

[44] BERTHOLD, Michael; HAND, David J. Intelligent data analysis: an introduction. Springer Science & Business Media, 2003.

[45] PAL, Nikhil R.; BEZDEK, James C. On cluster validity for the fuzzy c-means model. IEEE Transactions on Fuzzy systems, 1995, vol. 3, no 3, p. 370-379.

- [46] BALASKO, Balazs; ABONYI, Janos; FEIL, Balazs. Fuzzy clustering and data analysis toolbox. Department of Process Engineering, University of Veszprem, Veszprem, 2005.
- [47] DUSSÁN, Miguel. Sistema de recomendación web basado en la actividad de los usuarios de la Universidad Nacional de Colombia. Universidad Nacional de Colombia, Bogotá, Colombia. 2012.
- [48] AKSOY, Selim; HARALICK, Robert M. Feature normalization and likelihood-based similarity measures for image retrieval. Pattern recognition letters, 2001, vol. 22, no 5, p. 563-582.
- [49] MOHAMAD, Ismail Bin; USMAN, Dauda. Standardization and its effects on K-means clustering algorithm. Research Journal of Applied Sciences, Engineering and Technology, 2013, vol. 6, no 17, p. 3299-3303.
- [50] MARTINEZ A Francisco. Análisis de las series temporales de los precios del Mercado eléctrico mediante técnicas de clustering. Departamento de lenguajes y sistemas informáticos. Universidad de Sevilla. Doctorado en ingeniería informática.
- [51] BEZDEK, James C. Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, 1981.
- [52] DUNN, J. C. Indices of partition fuzziness and the detection of clusters in large data sets. Fuzzy Automata and Decision Processes, Elsevier, New York, 1977.
- [53] DAVIES, David L.; BOULDIN, Donald W. A cluster separation measure. IEEE transactions on pattern analysis and machine intelligence, 1979, no 2, p. 224-227.
- [54] CHANDRAKALA, S.; SEKHAR, C. Chandra. A density based method for multivariate time series clustering in kernel feature space. En Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, 2008. p. 1885-1890.
- [55] GUO, Chonghui; JIA, Hongfeng; ZHANG, Na. Time series clustering based on ICA for stock data analysis. En Wireless Communications, Networking and Mobile Computing, 2008. WiCOM'08. 4th International Conference on. IEEE, 2008. p. 1-4.
- [56] DING, Hui. Querying and mining of time series data: experimental comparison of representations and distance measures. Proceedings of the VLDB Endowment, 2008, vol. 1, no 2, p. 1542-1552.
- [57] LIAO, T. Warren. Clustering of time series data—a survey. Pattern recognition, 2005, vol. 38, no 11, p. 1857-1874.

- [58] CACERES, Gustavo.; RODRIGUEZ, Jorge E. Agrupamiento de datos de series de tiempo. Estado del arte, Revista Vínculos, Vol. 8, Número 1, 2011, Páginas 210-231.
- [59] IGLESIAS, Félix; KASTNER, Wolfgang. Analysis of similarity measures in times series clustering for the discovery of building energy patterns. Energies, 2013, vol. 6, no 2, p. 579-597.
- [60] D'URSO, Pierpaolo; MAHARAJ, Elizabeth Ann. Autocorrelation-based fuzzy clustering of time series. Fuzzy Sets and Systems, 2009, vol. 160, no 24, p. 3565-3589.
- [61] TORO, E.; GALLEGO, R.; ESCOBAR, A. Técnicas Metaheurísticas de Optimización. Universidad Tecnológica de Pereira, 2008.
- [62] CREG, Comisión de Regulación de Energía y Gas. Circular Creg 052 de 2010: Informe final tomo 2. Medidas para la reducción de pérdidas técnicas en los sistemas de Distribución. Colombia. Mayo 2010.
- [63] HOLLAND, John Henry. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence. MIT press, 1992.
- [64] WU, Jingwei; HASSAN, Ahmed E.; HOLT, Richard C. Comparison of clustering algorithms in the context of software evolution. En Software Maintenance, 2005. ICSM'05. Proceedings of the 21st IEEE International Conference on. IEEE, 2005. p. 525-535.
- [65] BITTENCOURT, Roberto Almeida; GUERRERO, Dalton Dario Serey. Comparison of graph clustering algorithms for recovering software architecture module views. En Software Maintenance and Reengineering, 2009. CSMR'09. 13th European Conference on. IEEE, 2009. p. 251-254.
- [66] GALLEGO, Ramón A.; ESCOBAR, Antonio H.; RODAS, Darío E. Algoritmo Genético Especializado para la Localización Óptima de Capacitores en Sistemas de Distribución de Energía Eléctrica. Scientia et technica, 2004, vol. 10, no 25.
- [67] KARYPIS, George; HAN, Eui-Hong; KUMAR, Vipin. Chameleon: Hierarchical clustering using dynamic modeling. Computer, 1999, vol. 32, no 8, p. 68-75.
- [68] SAVARESI, S. M.; BOLEY, D. L.; BITTANTI, S.; GAZZANIGA, G. Cluster selection in divisive clustering algorithms. En Proceedings of the 2002 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics, 2002. p. 299-314.

[69] JAIN, Anil K.; DUBES, Richard C. Algorithms for clustering data. Prentice-Hall, Inc., 1988.

[70] MARTINEZ DE PISÓN, F.; ORDIERES, J.; PERNÍA, A.; ALBA, F. Minería de datos en series temporales para la búsqueda de conocimiento oculto en históricos de procesos industriales. Actas del III Taller Nacional de Minería de Datos y Aprendizaje, TAMIDA, 2005, p. 31-38.

[71] EGAN, M. A.; KRISHNAMOORTHY, M.; RAJAN, K. Comparative study of a genetic fuzzy c-means algorithm and a validity guided fuzzy c-means algorithm for locating clusters in noisy data. En Evolutionary Computation Proceedings, 1998. IEEE World Congress on Computational Intelligence. The 1998 IEEE International Conference on. IEEE, 1998. p. 440-445.

7. ANEXO – PARAMETRIZACIÓN DEL AG

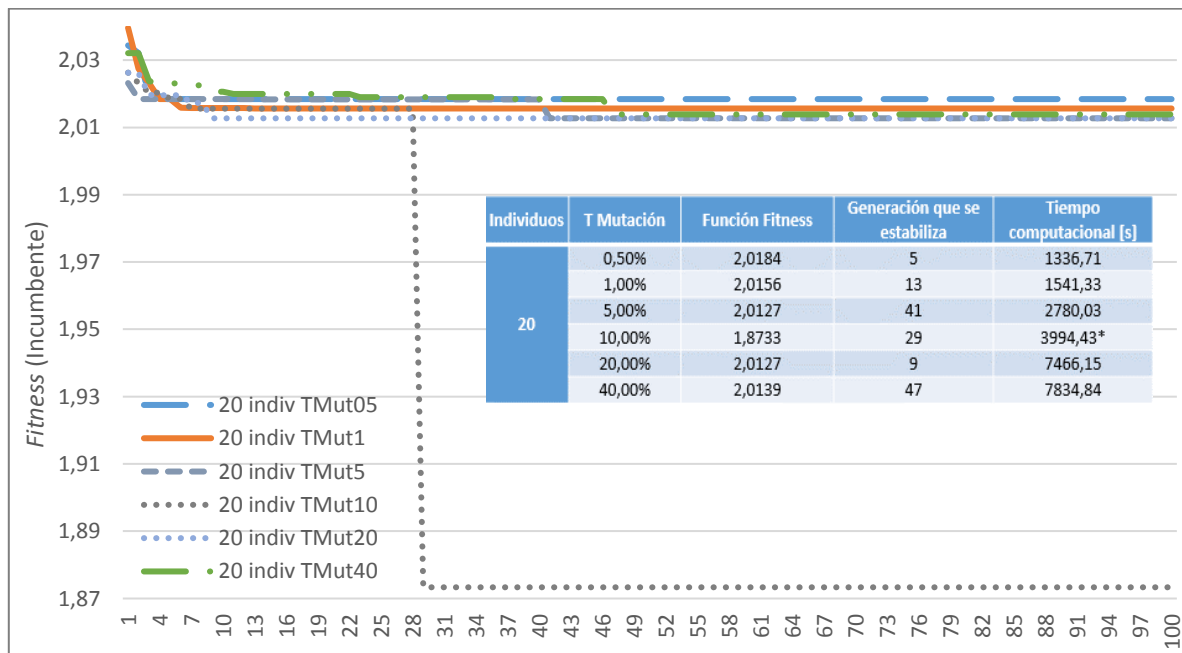
Las características básicas, así como la eficiencia de los algoritmos genéticos, dependen en gran medida de los parámetros de control²³ definidos y por el criterio de parada adoptado. En este anexo se determinan los parámetros de control a ser aplicados en el AG para garantizar soluciones de calidad en relación con la función *fitness* alcanzada en un tiempo computacional razonable.

7.1. POBLACIÓN INICIAL Y TASA DE MUTACIÓN

En las Figuras 32, 33 y 34 se evalúa la función *fitness* variando el tamaño de la población *I* y la tasa de mutación ρ_m . Los valores seleccionados para parametrizar el AG son aquellos para los cuales se alcance el menor valor de la función *fitness*, verificando que el tiempo de ejecución sea razonable.

- Población inicial *I*: 20
- Tasa de mutación ρ_m : Fija de 0,5%, 1%, 5%, 10%, 20% y 40%. Tipo *Multi point*
- Tasa de *crossover*: Fija de 0,9

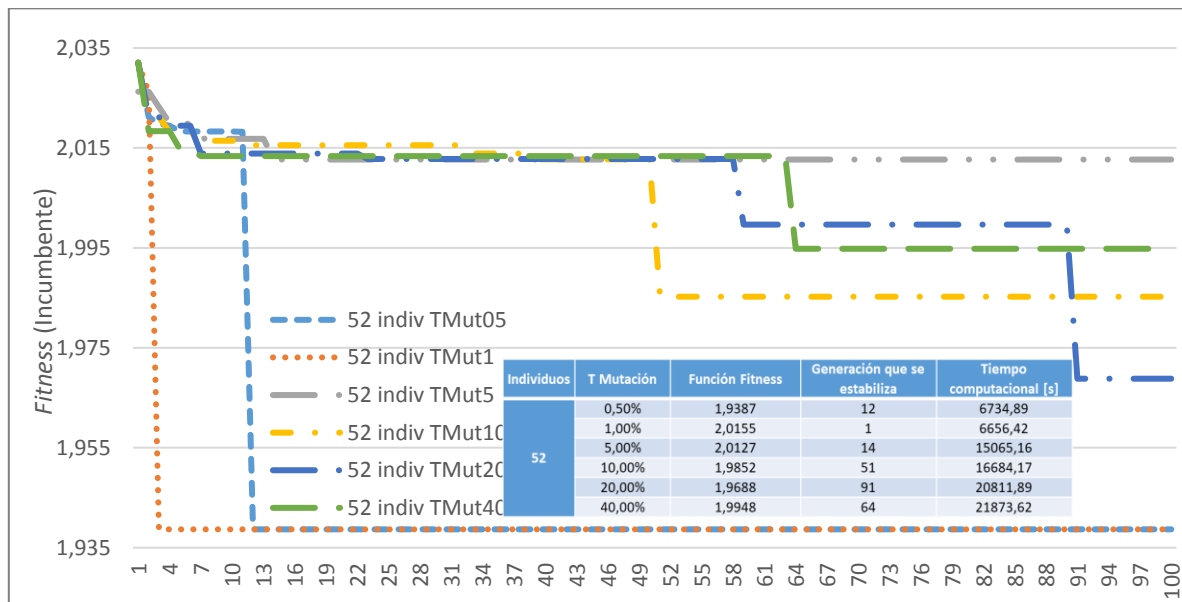
Figura 32. Desempeño AG 20 Individuos



²³ Tamaño de la población inicial, tasa de recombinación y tasa de mutación

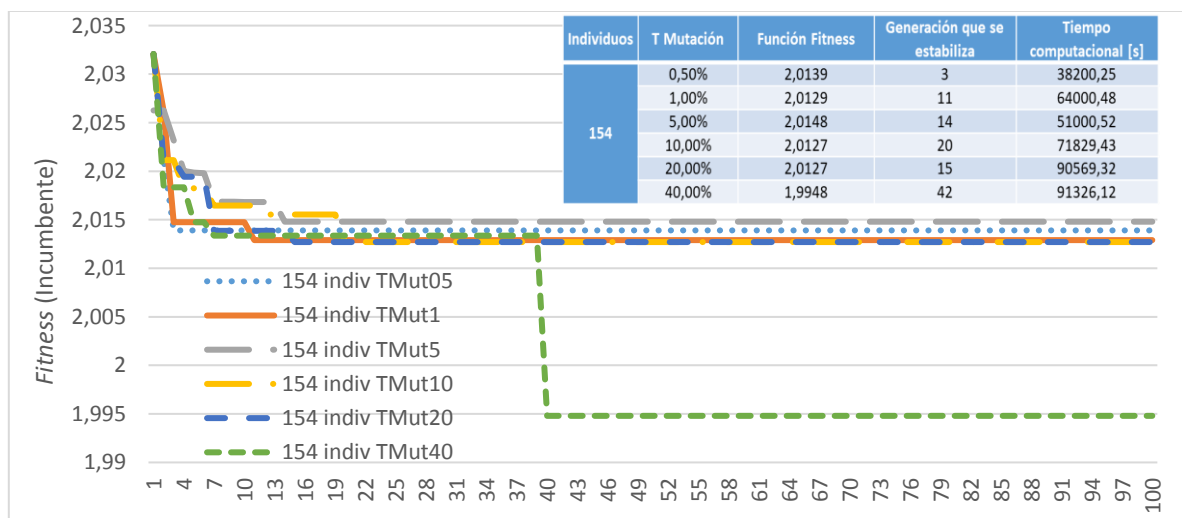
- Población inicial I: 52
- Tasa de mutación ρ_m : Fija de 0,5%, 1%, 5%, 10%, 20% y 40%. Tipo *Multi point*
- Tasa de *crossover*: Fija de 0,9

Figura 33. Desempeño AG 52 Individuos



- Población inicial I: 154
- Tasa de mutación ρ_m : Fija de 0,5%, 1%, 5%, 10%, 20% y 40%. Tipo *Multi point*
- Tasa de *crossover*: Fija de 0,9

Figura 34. Desempeño AG 154 Individuos



El menor valor de la función *fitness* se alcanza para una población de 20 individuos con una tasa de mutación de 10% (0.1).

7.2. TIPO DE MUTACIÓN Y TASA DE CROSSOVER

Con el tamaño de la población y la tasa de mutación definidos anteriormente, se evalúa el AG variando la tasa de *crossover* de 0,7 a 1 en pasos de 0,1 y variando el tipo de mutación entre *single point* y *multi point*. Los parámetros seleccionados de tasa de *crossover* y tipo de mutación son aquellos para los cuales el AG presente menor *función fitness*. En la Tabla 13 se muestran los resultados alcanzados.

Tabla 13. Parametrización Tasa de *Crossover* y tipo de mutación

Tasa <i>Crossover</i>	Tipo mutación	Función <i>Fitness</i>
0,7	<i>Single Point</i>	1,9852
	<i>Multi Point</i>	2,0139
0,8*	<i>Single Point</i>	2,0127
	<i>Multi Point*</i>	1,8733*
0,9	<i>Single Point</i>	2,0155
	<i>Multi Point</i>	1,9688
1	<i>Single Point</i>	2,0184
	<i>Multi Point</i>	1,9688

De la Tabla 13 se observa que el mejor ajuste se obtiene para la tasa de *crossover* de 0,8 tipo *Multi point*.

7.3. CRITERIO DE PARADA

El criterio de parada del AG es importante ya que hace parte de las características básicas de operación, así como también es determinante para la eficiencia del mismo. Con el fin de determinar un criterio de parada adecuado, se corre el AG para 1000 generaciones con los parámetros de control antes determinados. En cada corrida del AG se guarda la generación para la cual el algoritmo alcanza su valor mínimo de función *fitness*, perdiendo diversidad en las generaciones siguientes. En la Tabla 14 se muestran los resultados obtenidos:

- Población inicial I: 154
- Tasa de mutación ρ_m : Fija de 10%, tipo: *Multi point*
- Tasa de *crossover*: Fija de 0,8

Tabla 14. Parametrización criterio de parada

Simulación	Generación en que se estabiliza
1	47
2	29
3	41
4	51
5	14
6	20
7	42
8	30
9	29
10	43

Se puede concluir que a partir de la generación 50 se pierde diversidad genética y el algoritmo alcanza su mejor nivel, por lo cual 50 iteraciones es un buen dato como criterio de parada. Sin embargo, para darle oportunidad al algoritmo de evolucionar posteriormente y teniendo en cuenta que el proceso es corrido en modo *offline*, se selecciona un criterio de parada de 100 iteraciones.